技术白皮书

人工智能与图数据库技术

通过领域知识和关联数据提高AI性能

Amy E. Hodler, Mark Needham & Jake Graham

俞方桦 博士, 编译

技术no书

目 录

人工智能与图数据库技术

通过领域知识和关联数据提高AI性能

Amy E. Hodler, Mark Needham & Jake Graham

人工智能（AI ）的概念历史悠久。简单地给个定义，人工智能是一种解决方案或一套工具，可以模仿人类智能的方式解决问题。通常，它最实际的目标是进行预测：对事物进行分类（例如添加标签）或预测值（例如系列中预期的下一个数字）。

从更广泛的意义上说，AI有两类：狭义的和普遍的方法。狭义AI专 注于很好地执行某项任务，例如图像识别。更普遍的AI包括智能规划、自然语言理解、对象识别、机器学习或解决问题的多种能力。今天的人工智能解决方案虽然大多属于狭义的人工智能类别，但它们在适用于新情况方面拥有越来越广泛的能力，因此随着时间的推移也变得更加强大。

使AI应用程序能够具有更广泛的能力的一种方法是为它们提供领域知识(又称上下文 ，context)，为它们提供相关信息以用于解决手头的问题。

考虑一下自动驾驶汽车的情行。在雨天条件下自动驾驶车辆是很困难的，因为在多雨的条件下会有更多可变因素（想想太阳雨、暴风雨、冬天雨雪混合等不同天气类型，还有光线的因素、可能从左上或右上方照射下来）。

自动驾驶汽车的AI需要学习光线和天气条件的每种可能组合，但是事实上在所有可能的情况下对其进行训练上是不可能做到的。换一个角度，如果能为AI提供关联的领域相关知识（例如下雨的夜晚、夜晚和温 度的关联），则可以组合来自多个领域相关的信 息并推断出下一个要采取的操作（如减速或打开前灯）。

图和图数据库技术管理关联数据并定义关系。通过应用领域相关知识增强AI的性能，图技术提供了一种有效的技术手段来实现复杂AI应用程序的开发。

什么是人工 智能 (AI) ?

人工智能技术有三大类，每种类别都以不同的方式解决问题。 人工智能是一个总括性术语，包括机器学习（ML）和深度学习（DL）的各个子集。

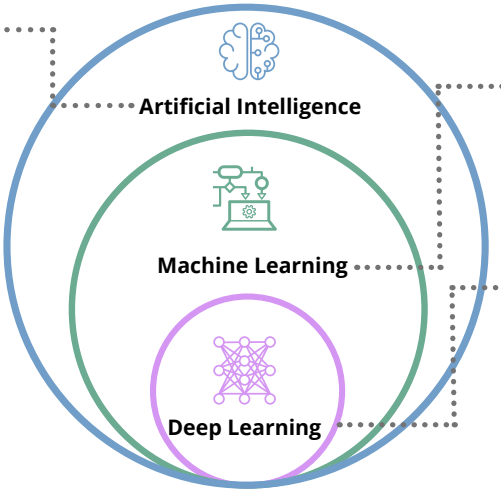
AI 是一类计算机过程，通过学习和模仿人类决策的方式解决问题。
请注意，这不需要具有实际智能。 然而，它确实为许多问题打开了大门，以执行人类智能所特有的任务。AI是解决方案的目标，机器学习本质上是实现它的一种方法。

机器学习 (ML) 使用算法帮助计算机学习特定任务的示例、实现渐进式改进，而无需显式编程。 “训练”AI涉及向算法提供大量数据，以使其能够学习如何处理这 些信息 。 机器学习的“学习”部分意味着相关算法通过迭代以优化目标函数，例如实现最小化误 差或损失。机器学习同时是动态的，能够在呈现更多数据时自行修正。

深度学习 (DL)使用以神经网络为模型的级联处理层来学习数据 表示，例如特征或分类。深度学习的“深层”部分是指多个隐藏的抽象层。这些图层实现了具有层次结构的特征集，例如向水果类别添加形状、大小和气味。

人工智能 (AI) 是一类计算机过程，通过学习和模仿人类决策的方式解决问题。

今天的AI解决方案主要用于非常具体的任务，而不是通用应用。



机器学习 (ML) 使用算法帮助计算机学习特定任务的示例、实现渐进式改进，而无需显式编程。

深度学习 (DL)

使用以神经网络为模型的级联处理层来学习数据表示，例如特征或分类。

人工智能由几个技术子集组成，每个技术子集都以不同的方式解决问题。

假设我们正试图解决一个现实世界的问题并做出一个决定，这个决定要求我们拥有正确的领域知识、并尝试以某种方式自动化或简化决策过程。

领域知识对AI的重要性

对于人类和人工智能而言，领域知识对决策至关重要。成年人每天做出成千上万的决定（有人说大约35,000个），而且大多数都取决于我们周围的环境或我们看待世界的角度。

如果我们正在安排旅行，会考虑旅行是为了工作、娱乐还是与他人同行，因此最后的决定有很大差异。在人类语言中，话语的实际含义高度依赖于情境、谁使用短语以及其语调。例如，如果一个人说“滚出去！”，其真实意思可能会表达一个友好的玩笑，也可能是真正生气要求别人离开房间。

人类使用情境学习来确定在某种情况下什么是重要的、以及如何将其应用于新情况。如果要求人工智能来做出更接近人类的决策，则需要借助大量领域知识。如果没有外围设备和相关信息，AI需要更详尽的训练、更多的规范性规则和更具体的应用。

图可以实现领域相关知识的四种方法

至少在四个主要区域，图可以为AI提供领域相关知识。我们将在本白皮书的其余部分详细介绍以下各节。

首先是知识图谱，它为决策支持提供领域相关知识/上下文（例如，为呼叫中心员工或现场支持工程师），并且帮助确保答案适合于该特定情况（例如，在多雨驾驶条件下的自动车辆。

其次，图提供更高的处理效率，因此借助图来优化模型并加速学习过程可以有效地增强器学习的效率。

第三种，基于数据关系的特征提取分析可以识别数据中最具预测性的元素。基于数据中发现的强特征所建立的预测模型拥有更高的准确性。

第四种，也是最后一种，图提供了一种为AI决策提供透明度的方法，这使得通过AI得到的结论更加具有可解释性。

假设我们正试图解决一个现实世界的问题并做出一个决定 ，这个决定要求我们拥有正确的领域知识 、并尝试以某种方式自动化或简化决策过程。

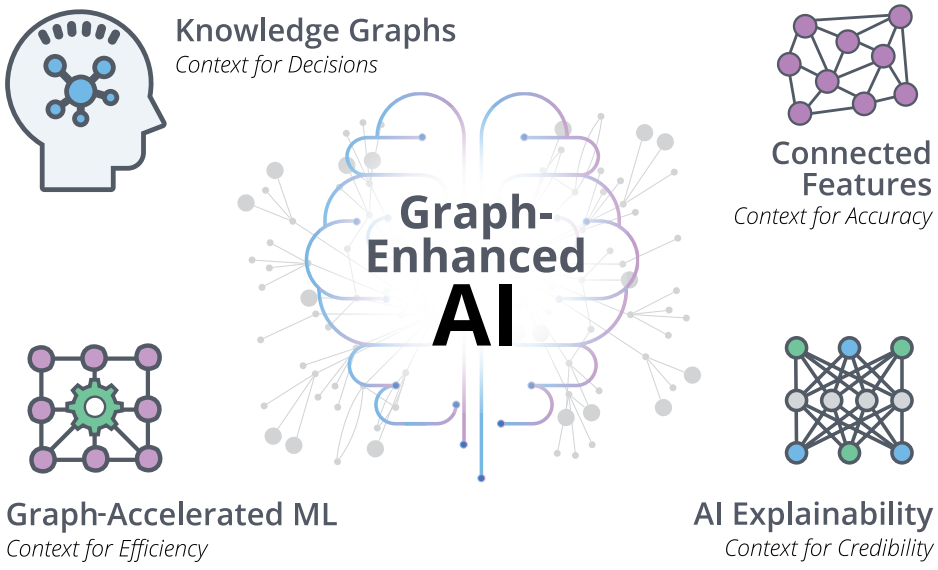
知识图谱j决策的背景

最快进入实际应用的AI领域之一是决策支持。假设我们正试图解决一个现实世界的问题：做一个决定，这个决定要求我们拥有正确的领域知识，并尝试以某种方式自动化或简化该过程。

知识图谱提供了一种简化工作流程、自动化响应过程和扩展智能决策的方法 。在高层次上，知识图谱是相互关联的事实集合，以人类可理解的形式描述现实世界的实体、事实或事物，及其相互关系。与具有平面结构和静态内容的简单知识库不同，知识图谱通过获取和集成相邻的信息以获得新知识。

以下是知识图谱的一些关键特征j:

- 知识图谱需要围绕相关属性进行连接。由于并非所有数据都是知识，我们寻找的是与领域相关的信息。
- 知识图谱是动态的，图本身可以理解连接实体的内容，无需手动为每条新信 息编写程序。知识图谱能够把那些对我们重要的属性进行适当的关联，基于我们已经对它们建立的关系。
- 知识图谱需要能够被理解。有时我们说它是有语义的，因为知识本身告诉我们 是什么。智能元数据帮助我们遍历图以查找特定问题的答案，即使一开始我们并不明确地知道如何要求它做到。
- 实际上，知识图谱通常包含异构数据类型。它结合并揭示了信息孤岛之间的联系。



知识图谱根据知识类型和使用的数据分成相关的三个类别

领域知识丰富的知识图谱

为内部知识文档和文件创建知识图谱，并赋予元数据标记。

应用案例:

- 搜索引擎
- 客户服务
- 文档分类

当今市场上有三种主要的知识图谱类别领域:知识图谱、外部感知图谱和 自然语言处理（NLP）图谱。

领域知识丰富的知识图谱

为内部知识文档和文件创建知识图谱，并赋予元数据标记。

外部感知知识图谱

为外部数据源建立知识图谱，并汇总和映射到感兴趣的实体。

自然语言知识图谱

为技术术语、首字母缩写词、拼写错误等建立知识图谱。

领域相关知识图谱

领域知识丰富的知识图谱解决了这样的挑战j简单的基于关键字的文档搜索、或简单识别单个单词的重要性对于检索拥有大量异构数据的知识库并不准确和有效。

建立知识图谱则使我们能够将内部文档、文档相关的领域知识、以及元数据标记相结合，在图数据库中能够更快地连接和遍历这些知识。

对于领域知识丰富的知识图谱，最常见的用例是Google的搜索引擎，另外文档分类和客户支持也是常见的应用场景。例如，如果我们能够根据每年收到的数以万计的复杂技术支持问题，向技术支持人员快速展示最类似的问题、以及如何解决问题的方法和相关文档，那么会大大加快问题解决的速度。

包含丰富领域相关知识的知识图谱适用于以文档形式获得和保存大量知识的组织。知识图谱有助于填补信息收集、与能够查找和应用该信息（通过数据关联）之间的差距。一个成功的例子是美国宇航局的经验教训数据库，该数据库记录了过去5年来的任务和项目知识。

外部感知知识图谱

外部感知知识图谱聚合外部数据源并将它们映射到感兴趣的内部实体。例如，在评估供应链风险时，我们可能希望查看所有供应商、他们在所有地方的工厂、以及我们所有的供应线，以分析中断风险。 另外，还可以考虑当特定地点发生自然灾害时会如何影响供应链并识别在那些地点附近的类似供应商。

一般而言，我们需要能够从市场中收集大量信息j感知信息，确定与领域相关的信息内容，并将其呈现给需要的人。除了供应链监控之外，外部洞察感知还用于分析合规风险、市场活动的影响和识别销售机会。

例如，路透社（现更名为Refinitiv）拥有关于企业财务内容的知识图谱，使组织能够连接外部和内部知识，并在大市有时间做出反应之前快速做出最佳财务决策。

外部感知知识图谱

外部数据源聚合映射到感兴趣的实体。

应用案例

- 供应链合规风险
- 市场活动
- 销售机会

THOMSON REUTERS

pitney bowes

自然语言处理知识图谱

建立特定技术术语、产品名称、行业首字母缩略词 、部件号甚至常见的拼写错误的知识图谱。

应用案例

- 增强型搜索
- 聊天机器人
- 改进的内容分类

ebay

CAT

自然语言处理知识图谱

自然语言处理（NLP）知识图谱包含关于人类语言的复杂性和细微差 别的知识。 NLP知识图谱需要了解公司的特定技术术语、产品名称、行业首字母缩略词、部件号甚至常见的 拼写错误。这是分析师创建知识图谱以映射含义和构建本体的地方，在此基础上进一步改进搜索并提供更相关的结果。

重型设备制造商Caterpillar使用NLP知识图谱来支持自然语言搜索，并从数千份保修文档中提取含义。另一个例子是eBay App for Google Assistant，它使用所有三种类型的知识图谱：相关领域知识、外部感知和NLP，来引导购物者获得完美的产品。

今天，基于图的AI应用程序的许多实现都利用了知识图谱。本文的其余部分会继续探讨图技术对提升AI应用有重要贡献的其他领域。

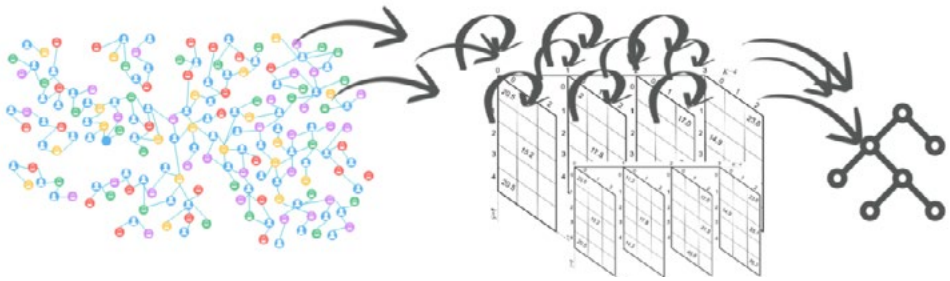
知识图谱增强机器学习：领域知识提高运行效率

当前的机器学习方法通常依赖于存储在表中的数据。机器学习这些数据充其量只是资源密集型操作。超过一半的受访企业CIO表示，迭代模型的学习效率是将AI项目从概念转化为生产力的最大挑战之一。

知识图谱提供了提高效率的领域相关内容，因为它将数据连接起来，从而在关系上实现了多个分离度、有利于大规模快速遍历和分析。从这一意义上说，**图加速了机器学习的效果**。

人类天生具备自然地连接相关信息的能力。举个例子，考虑一下人们在被问到“这张关于狗的照片让你想起什么？”时的反应。人类不需要运行繁复的程序，例如最近邻分类器，将该狗与所有像狗的对象进行比较。我们几乎立即就会发现它哺乳动物 – 而不是人类或其他无生命的物体 – 并将它归类为狗

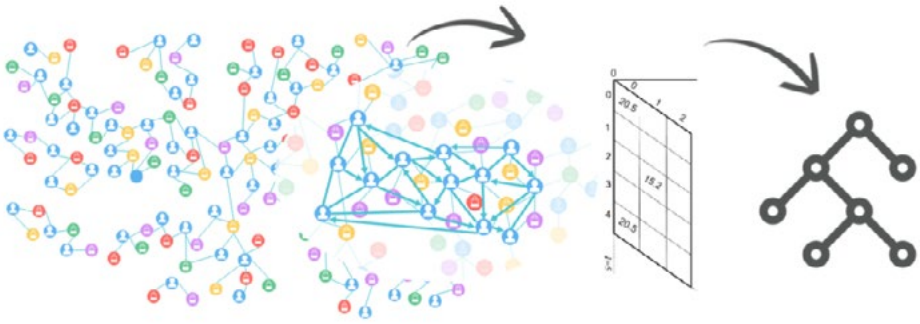
当数据以表的形式存储时，需要多次迭代才能连接它。例如，当过滤过程需要在数据管道中连接多个表以揭示某种关系时，其效率变得非常之 低。而像协同过滤等数据科学实践却往往需要频繁连接(JOIN)、索引和查找多个表来得到要求的结果。



在许多机器学习系统中，来自现实世界关联的数据被存储为表，然后通过迭代连接以产生决策树。

扩展性是关于机器学习效率的另一 个问题。机器学习算法可能需要针对所有数 据进行计算。 为避免这种情况，分析师会 手动创建数据子集。 然而，这些方法往往会减慢迭代速度，因为它们要么计算 密集、要么需要人为参与。 简单的图查询则可以通过返回仅包含所需数据的子图来加速此过程。

知识图谱提供了提高效率的领域相关内容，因为它将数据连接起来，从而在关系上实现了多个分离度、有利于大规模快速遍历和分析。



对存储在图数据库中的关联数据进行训练效率更高。
图过滤过程相当地高效特别是与典型的手动设置或统计推断相比。

使用图，我们可以快速提取可以用于预测的特征、并重新转换数据以便在机器学习的管 道中使用。 例如，从图中我们将相关的数据子集（例如强连通子图 ）提取为表状格式以进行模型构建。

关系特征: 领 域知识提高准确度

关系往往是行为的最强预测因素。

例如，研究表明，你更大的朋友圈在预测你是否会投票方面比你的直接朋友圈会是一个更好的指标（在这种情况下，朋友的朋友比直接的朋 友更有影响力）。关联数据的特征可以从图中与连接相关的指标得到，例如进出节点的关系数量(入度和出度)、潜在三角形或共同邻居的数量。

当前的机器学习方法通常依赖于从表构建输入数据 。这意味着尝试抽象、简化并且(有时)则完全遗漏了大量的预测关系和领域相关知识。通过将关联的数据和关系存储为图，可以更直接地提取关系的特征、并更轻松地包含所有重要信息。

关系特性可以在许多行业中使用，尤其在调查欺诈和洗钱等金融犯罪领域。在 这 些 案例中，犯罪分子经常试图通 过多层次混淆和复杂网络关系来隐藏犯罪活动 。传统方法可能无法检测到这种行为，而这 正是图在提取关系特征方面所擅长的领域。

使用关系特征有几种不同的方法。 在上一节我们介绍 了通过特征提取重新格式化我们的数据； 具有更大价值的则是特征工程(Feature Engineering)，它结合并处理数据以创建新的、更有意义的特征。

特征工程可能需要同时涉及简单查询或图算法。当我们确切地知道想要找到什么时，例如确定某人的网络中有多少已知的欺诈者，特定的查询就可以很好地解决问题。

像Louvain算法这样的图社区检测算法可识别紧密的社区和关系层次结构，
以创建用于改进机器学习预测效果的带权重的特征。

“我们越来越多地了解到，从朋友和朋友的朋友那里获取关于某人的信息，要比自己手头掌握的关于这个人信息在进行预测方面能 获得更好的结果。”

- James Fowler, Connected

但是，我们应该使用图算法来寻找我们已经知道和想要得到的一般结构特征、而不是确切的模式。例如，图算法简化了发现类似紧密社区的异常情况的过程，这些社区可能是欺诈团伙或 洗钱网络。然后，我们可以在紧密社区中对节点进行评分，并提取该信息以用于训练机器学习模型。

最后，我们使用图算法进行特征选择，以将模型中使用的特征数量减少到一个最相关的子集。一个例子是可以使用像PageRank这样的算法来查找影响最大的特征，然后可以用来确定哪些属性最 能预测欺诈。这有助于消除不太重要的特征带来的噪音，并减少过度拟合(即模型在其训练数据上被过度训练)。

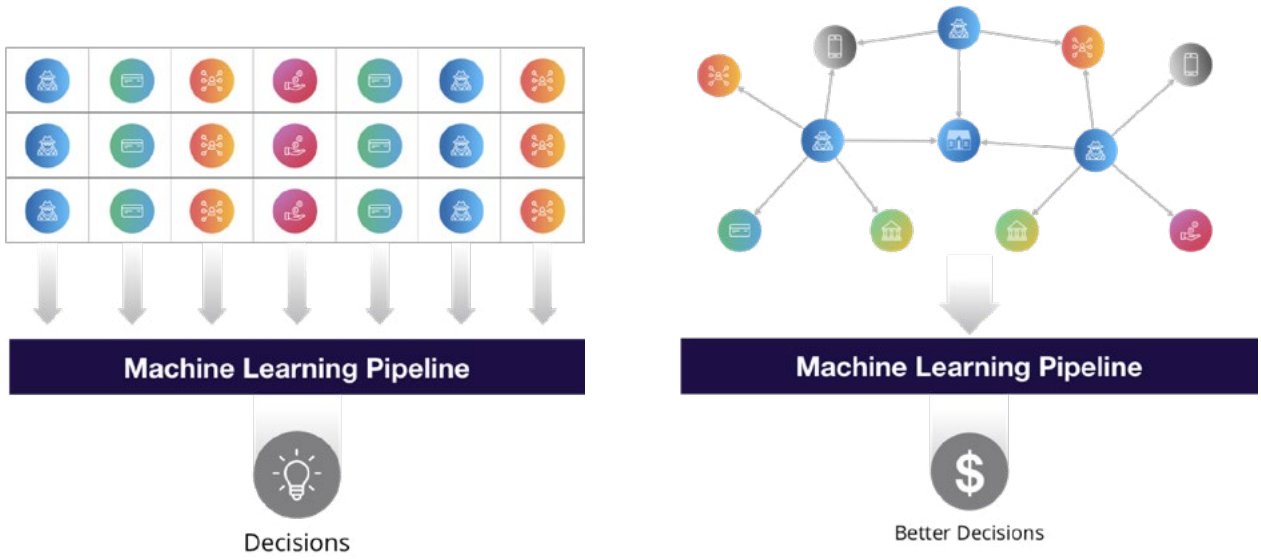
使用关系特征可以最大限度地提高模型的预测能力，同时扩展解决方案的应用范围。

可解释的人工智能: 领域知识提供可信度

在应用AI的过程中一个最大挑战是理解AI究竟是何做出特定决策的。 可解释的人工智能是一个还在逐步形成的领域。然而已经有相当多的研究表明，图使人工智能预测更易于追踪和解释。

这种能力对于AI的长期应用至关重要，因为在许多行业，例如医疗保健、信用风险评估和刑事司法，我们必须能够解释AI如何以及为何做出决策。 这是图可以应用领域相关知识来提高可信度的地方。

有许多机器学习和深度学习的例子实际上提供了错误的答案。 分类器可能产生导致错误分类的关联，例如将狗分类为狼。然而， 理解导致AI解决方案做出特定决定的原因有时是一项重大挑战。

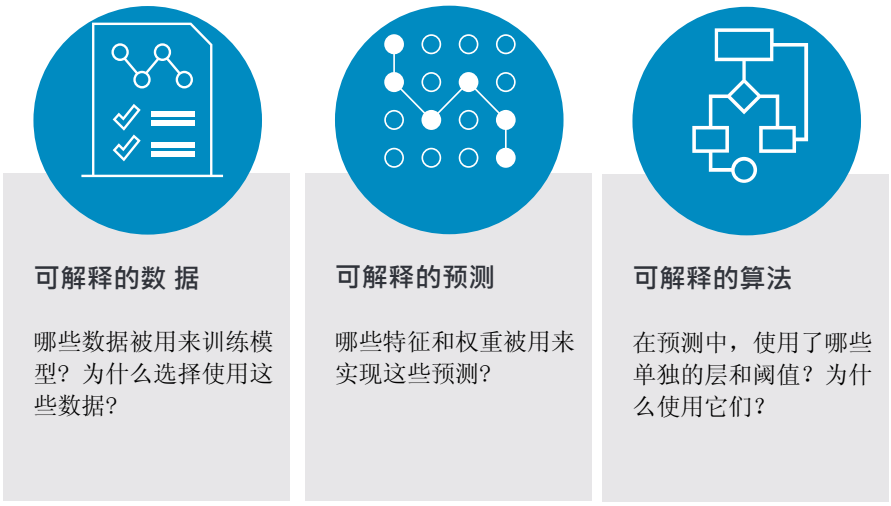


基于“平面数据”的传统方法简化或完全省略了可用于预测的 关系。(包含数据和关系的)图为模型添加了高度预测性功能，在不改变当前工作流程的情况下提高了准确性。

特别提示

Neo4j图数据库平台包含超过35种图算法，用于探索和分析关联的数据。你可以在我们的网站上注册并获得一本O’Reilly出版的、关于图算法的免费电子书。

有三类可解释性与我们提出的问题类型有关：



关于AI可解释性的问题有助于我们理解数据、预. 测和算法影响决策。

可解释的数据意味着我们知道用什么数据来训练我们的模型、以及为什么。不幸的是，这并不像我们想象的那么简单。如果考虑大型云服务提供商，或者诸如Facebook等拥有大量数据的公司，通常很难知道用于其算法的确切数据。

领域相关知识图谱通过数据关联相当容易地解决了数据的可解释性问题，因而被当今大多数顶级金融机构所采用。建立知识图谱需要将数据存储为图的结构，这使得跟踪数据的更改历史、数据在哪些地方被使用、以及谁使用了哪些数据变得非常简单。

另一个具有巨大潜力的领域是对**可解释预测**的研究。 这意味着可以知道特定预测使用了哪些特征、以及什么权重。当前，在使用图进行可解释的预测方面有很多活跃的研究。

例如，如果我们将神经网络中的节点与带标记的知识图谱关联起来，当神经网络使用到一个节点时，我们可以很快根据知识图谱获得所有关联节点的相关数据。这样，我们可以遍历激活的节点、并从其邻居数据推断出有意义的解释。

最后，**可解释的算法**使我们能够了解是哪些单独的层和阈值选择产生了相关预测。 在这个领域形成实际解决方案还有很长的路要走，但前景诱人。一些研究包括在加权图中构造张量线性关系(tensor)。初期的成果显示我们确实有可能在每一层找到特定的解释和相关系数。

Neo4j图数据库平 台使您能够更快地构建智能应用程序、有能力寻找到新市场机会、满足客户需求、提高业务产出并应对最紧迫的业务挑战。

结 论

在本白皮书中，我们考虑了知识图谱为人工智能添加领域相关知识的四种方式知识图谱为AI提供领域知识、领域知识提升ML的效率、领域知识提高特征提取的准确性、以及领域知识提供AI可解释性和可信度。

AI和机器学习具有很大的应用潜力，而图解锁了这种潜力。这是因为图数据库技术支持领域相关知识和关联数据、使AI变得更广泛适用。

如果您正在构建AI解决方案，请考虑使用图和图数据库技术来为其提供领域相关知识。Neo4j图数据库平台使您能够更快地构建智能应用程序、有能力寻找到新市场机会、满足客户需求、提高业务产出并应对最紧迫的业务挑战。

想更多了解 Neo4j?

欢迎随时 和 我们联 系:
apac@neo4j.com
neo4j.com/contact-us