arXiv:2303.02166v1 [cs.DB] 3 Mar 2023

迈向图机器学习增强的知识图谱平台

Hussein Abdallah康考迪亚大学,
hussein.abdallah@concordia.ca

Essam Mansour康考迪亚大学,
essam.mansour@concordia.ca

摘要——**本文提出了 KGNet：一种按需提供的图机器学习（GML）即服务，构建于 RDF 引擎之上，用于支持图机器学习增强的 SPARQL 查询。KGNet 通过识别任务特定子图，自动化在知识图谱上训练图机器学习模型的过程。这有助于减少与任务无关的知识图谱结构和属性，从而提升可扩展性和准确率。在 KG 上训练图机器学习模型时，KGNet 会将训练后的模型元数据收集为 RDF 图，称为 KGMeta；该图与 KG 中的相关子图相互链接。最后，所有训练后的模型都可通过类 SPARQL 查询访问。我们将其称为图机器学习增强查询，并将其记为 SPARQL^{ML}。KGNet 在现有 RDF 引擎之上支持 SPARQL^{ML}，作为使用图机器学习模型对知识图谱进行查询和推理的接口。KGNet 的开发在多个领域带来了研究机会，包括用于识别任务特定子图的元采样、在有限时间和内存预算等计算约束下实现图机器学习流水线自动化，以及 SPARQL^{ML} 查询优化。KGNet 支持不同的图机器学习任务，例如节点分类、链接预测和语义实体匹配。我们使用来自不同应用领域的两个真实知识图谱对 KGNet 进行了评估。与在整个知识图谱上训练相比，KGNet 在保持准确率相当或有所提升的同时，显著减少了训练时间和内存使用量。KGNet 源代码[1] 可供进一步研究。**

I. 引言

知识图谱（KG）基于从异构数据集中提取的语义构建而成，构建过程中使用了表示学习和分类模型等多种人工智能技术 [1]。图机器学习（GML）技术，例如图表示学习和图神经网络（GNN），是强大的工具，广泛用于将现实世界问题定义为知识图谱上的预测任务并加以解决。例如，推荐 [2] 和实体对齐 [3], 等问题中的节点分类任务可以使用图机器学习技术解决。同样，药物发现 [4]和欺诈检测 [5], [6]问题也可借助图机器学习技术作为链接预测任务来处理。

数据科学家通常使用知识图谱，而知识图谱通常存储在资源描述框架引擎中。他们负责开发
使用图机器学习流水线和 PyG [7] 、DGL [8], 等框架在这些知识图谱上训练模型。然而，图机器学习框架与资源描述框架引擎之间往往存在脱节。因此，在传统的图机器学习流水线中，首先需要将整个知识图谱从 RDF 三元组格式转换为邻接矩阵。之后，数据科学家需要从众多知识图谱嵌入（KGE）或图神经网络方法 [9], [10] 中选择合适的图机器学习方法来训练模型。对于普通用户而言，这项工作既耗时又费力。

[1]https://github.com/CoDS-GCS/KGNET

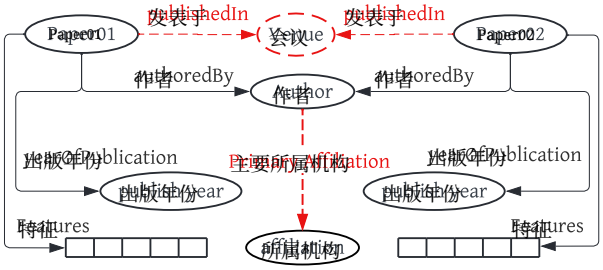


图 1：一个知识图谱，红色节点和边可由分类和链接预测模型即时预测。

```
1  prefix dblp: <https://www.dblp.org/>
2  prefix kgnet: <https://www.kgnet.com/>
3  select ?title ?venue
4  where {
5  ?paper a dblp:Publication.
6  ?paper dblp:title ?title.
7  ?paper ?NodeClassifier ?venue.
8  ?NodeClassifier a kgnet:NodeClassifier.
9  ?NodeClassifier kgnet:TargetNode dblp:Publication.
10 ?NodeClassifier kgnet:NodeLabel dblp:venue.}
```

图 2：SPARQL$^{ML}_{pv}$：SPARQL^{ML} 查询通过对图 1 所示知识图谱进行查询和推理，使用节点分类模型预测论文的会议。

此外，训练后的模型与存储知识图谱的资源描述框架引擎相互隔离。因此，自动在知识图谱上训练图机器学习模型，并通过类 SPARQL 查询访问训练后的模型至关重要。我们将这种查询称为 aSPARQL^{ML} 查询。

图 1 所示的知识图谱包含 DBLP 中已发表论文的信息 [11]。然而，传统 SPARQL 查询语言无法用于在知识图谱上应用图机器学习模型，例如预测节点类别或作者缺失的隶属关系链接。例如，图 1 中的会议节点是一个虚拟节点，可以使用节点分类（NC）模型进行预测。通过类 SPARQL 查询使用图机器学习模型执行 NC 查询此知识图谱，从而获取论文—会议节点，如 SPARQL$^{ML}_{pv}$ 中图 2 所示，十分有趣。该查询使用 *kgnet:NodeClassifier* 类型的模型，为每篇论文预测一个会议。第 8 至 10 行中的 SPARQL^{ML} 三元组模式将检索所有能够预测 dblp:venue 类型类别的 *kgnet:NodeClassifier* 类型模型。在三元组模式 ⟨?paper, ?NodeClassifier, ?venue⟩中，我们将*?NodeClassifier* 视为用户定义谓词。

实现图 2 所示的 SPARQL$^{ML}_{pv}$ 等查询面临多项挑战，包括：(*i*) 自动训练适用于各种任务的图机器学习模型，(*ii*) 基于准确率和推理时间优化用于图机器学习模型选择的 SPARQL^{ML} 查询，以及 (*iii*) 在查询执行期间与选定模型进行高效交互。此外，还需要将图机器学习模型无缝集成到资源描述框架引擎中。因此，用户应能遵循 SPARQL 模式匹配逻辑轻松编写 SPARQL^{ML} 查询，而无需显式使用用户定义函数（UDF）。

将图机器学习与 Neo4j [12] 或 Stardog [13]等现有图数据库集成的做法正日益普及。然而，尽管这些数据库提供了一些机器学习原语方法，例如使用 *Cypher* 语言实现 PageRank 和最短路径，但它们并未解决将图机器学习模型与资源描述框架引擎集成所面临的挑战。例如，Neo4j 图数据科学 [14] 在测试版中支持有限的图嵌入方法，例如 FastRP [15],Node2Vec [16], 和 Graph-SAGE [17]。然而，用户必须先单独训练这些模型。为应对这些挑战，我们需要将图机器学习引入存储在资源描述框架引擎中的数据，而不是将数据导入机器学习流水线。这样可以推动开发由 SPARQL 表达能力赋能的知识图谱数据科学库，从而基于知识图谱结构和语义实现更深入的分析与洞见发现。这些库将为数据科学家提供全面的知识图谱机器学习服务，服务于存储在资源描述框架引擎中的知识图谱。

本文提出了 KGNet，这是一种构建于 RDF 引擎之上、按需提供的图机器学习即服务，用于支持 SPARQL^{ML}查询，如图 3 所示。KGNet 通过两个主要组件扩展了现有 RDF 引擎：图机器学习即服务（GMLaaS）和 SPARQL^{ML} 即服务。KGNet 会针对节点分类或链接预测等任务，在知识图谱上自动训练图机器学习模型，并将训练模型的元数据维护为一个名为*KGMeta*的 RDF 图。为了缩短训练时间、减少内存使用量，同时提高特定任务 $\mathcal{A}$上的准确率，KGNet 会执行元采样，从较大的知识图谱中识别出与 $\mathcal{A}$相关且保留其基本特征的任务特定子图 KG' 。这使 KGNet 能够扩展至大型知识图谱。GMLaaS 负责：(*i*) 根据给定的时间或内存限制，选择近乎最优的图机器学习方法，以使用 KG' 训练 $\mathcal{A}$

预算，以及（*ii*）通过 HTTP调用与 RDF 引擎通信，请求对特定训练模型进行推理；（*iii*）存储与知识图谱相关的训练后的模型和嵌入。SPARQL^{ML} 服务会透明地执行以下操作：（*i*）维护 KGMeta，并将其与相关知识图谱相互链接；（*ii*）针对用户定义谓词优化图机器学习模型选择；（*iii*）最终将 SPARQL^{ML} 查询重写为 SPARQL 查询。

总而言之，本文的贡献包括 re:

- 在现有资源描述框架引擎之上构建了功能完备的支持图机器学习的知识图谱平台[2] 。
- 图机器学习即服务，根据给定的内存或时间预算自动训练图机器学习模型。

[2]https://github.com/CoDS-GCS/KGNET

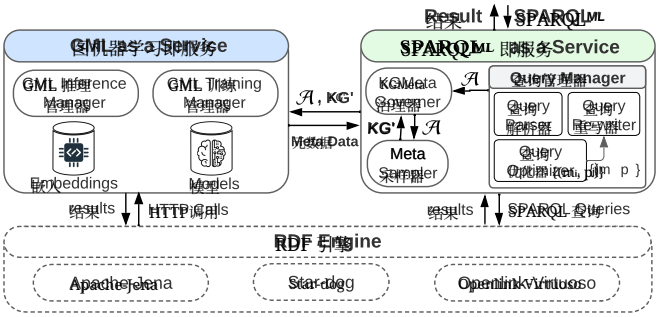


图 3：KGNet架构。该架构提供了一种接口语言（SPARQL^{ML}），使人工智能应用和数据科学家能够自动在知识图谱上训练图机器学习模型，从而基于训练好的模型查询知识图谱并进行推理。

这种自动训练利用通过我们的元采样方法提取的任务特定子图。

- SPARQL机器学习 即服务，用于执行元采样、在KGMeta中维护训练元数据，并优化图机器学习模型选择，即根据准确率和推理时间方面的约束，选择近似最优模型。
- 在真实知识图谱上，使用三种图机器学习任务和多种图机器学习方法进行了全面评估。实验表明，与在整个知识图谱上训练相比，KGNet取得了相当或更高的准确率，同时显著降低了训练时间和内存使用量。

本文其余部分的组织如下：第 II 节介绍现有图机器学习流水线的背景；第 III 节概述开发图机器学习赋能的知识图谱引擎所面临的主要研究挑战；第 IV 节介绍 KGNet 平台；第 V 节讨论对我们自动化图机器学习模型训练流水线的评估结果；第 VI 节和第 VII 节分别介绍相关工作和总结。

II. 背景：知识图谱的机器学习流水线 s

为知识图谱上的模型训练而开发的机器学习流水线可分为三大类：(*i*) 对表格格式的知识图谱数据进行传统机器学习，(*ii*) 对知识图谱嵌入进行传统机器学习，以及 (*iii*) 直接在知识图谱上训练图神经网络（GNN）。在使用表格格式知识图谱数据的传统机器学习方法中，知识图谱中的数据会被转换为内存中的数据帧，并使用特征工程技术和 Scikit-Learn 或 SparkMLib 等库训练经典机器学习分类器。相比之下，对知识图谱嵌入进行传统机器学习无需特征工程，而是为节点和边生成嵌入。Apple Saga [18]是这种方法的一个示例，它使用 DGL-KE [8]等图机器学习库来生成知识图谱嵌入。数据科学家可以灵活选择用于训练的机器学习方法。

近年来，图神经网络广受欢迎。因此，数据科学家经常使用它们来执行图机器学习任务。开放图基准（OGB） [19]对图神经网络训练流水线进行了标准化，强调了应对图机器学习任务和构建图神经网络训练流水线的最佳实践。图 4 总结了这一流水线，其中包括对

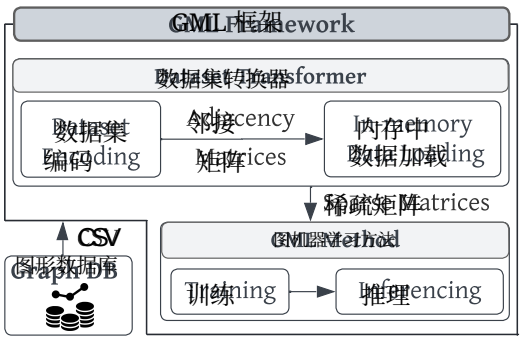


图 4：使用图机器学习框架的传统图机器学习流水线 [19] 。该流水线首先提取图数据，接着将数据转换为稀疏矩阵，以训练用于图机器学习任务的模型。最后，推理步骤可独立于图数据库进行结果预测。

知识图谱节点和边进行编码、生成邻接矩阵、将其加载到内存中，并使用特定方法训练图神经网络。

多种图机器学习框架（例如 DGL [8] 和 PyG[7],）提供了图神经网络方法的多种实现。这些框架通过将图加载到内存中并将其表示为图数据结构，再应用各种转换来支持数据转换。然而，对于大型知识图谱，现有图机器学习框架需要大量内存和处理时间，而且还要求用户深入了解各种图神经网络方法。相比之下，OGB 流水线较为简单，但这是一个半自动化流程，需要人工干预和机器学习专业知识，才能构建有效的流水线并选择合适的图神经网络方法。数据科学家可以根据时间或内存限制等各种约束，选择最合适的图神经网络方法。此外，如图 4 所示，训练后的模型与知识图谱引擎相互分离，这增加了数据科学家在推理知识图谱时应用模型的复杂性。

III. 图机器学习赋能的知识图谱引擎面临的挑战

本节重点介绍开发图机器学习赋能的知识图谱引擎所带来的开放性研究挑战与机遇。

A. 自动训练：方法选择与元采样

用于训练图机器学习任务模型的方法众多，如图 5 所示。这些方法主要可分为两类：知识图谱嵌入（KGE）方法和图神经网络（GNN）方法。知识图谱嵌入方法包括 TransE、RotatE、ComplEx 和 DistMult [10]。一些图神经网络方法支持在全图上进行采样，例如 Graph-SAINT [20], 、Shadow-SAINT [21], 和 MorsE [22]。关系图卷积网络（RGCN） [23] 和 GAT [24] 则是图神经网络全批次训练（不进行采样）方法的示例。我们的分类法还包含更多类别，如图 5 所示。

图机器学习方法在准确率、训练时间和内存需求方面差异显著。此外，图机器学习任务的复杂度可能因多种因素而异，例如知识图谱的规模以及与任务相关的节点/边类型数量。例如，链接预测可能比节点分类更消耗资源。不同

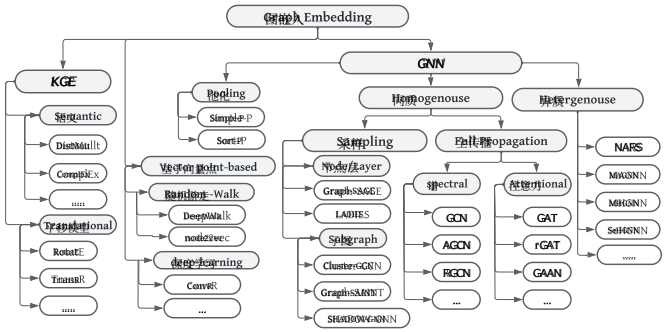


图 5：训练图机器学习模型的方法分类。

图机器学习方法在相同预算限制下的表现可能各不相同，而选择最佳方法取决于多个因素。因此，基于用户的时间和内存预算，为特定图机器学习任务自动构建训练流水线颇具挑战。例如，一些图机器学习方法采用全批次训练，需要更多内存预算。这些方法需要大量内存才能在大型知识图谱上训练模型。其他一些图神经网络方法可能会受到过平滑的影响，从而导致准确率下降。基于采样的图神经网络（小批次训练）方法采用不同类型的采样，这些采样在避免上述局限性方面各有不同。因此，根据给定的时间或内存预算，为特定任务自动选择图机器学习方法颇具挑战。

真实知识图谱可能包含数百万至数十亿个三元组，例如 DBLP [11] 和 MAG [25]。然而，在这些大型知识图谱上训练图机器学习模型需要庞大的计算资源，超出单台机器的能力。因此，需要从知识图谱中识别出一个规模较小且针对当前任务的训练数据集。这个过程称为元采样。元采样已被应用于多个应用领域，包括计算机视觉 [26], [27] 和语音识别 [28], , 用于提取针对给定任务的训练数据集。在图机器学习领域，元采样提供了一种机会：通过选择与任务相关的代表性子图，优化在大型知识图谱上的模型训练。这种方法有助于降低时间和内存需求，同时不牺牲准确率。因此，探索在训练图机器学习模型时利用元采样提取任务特定子图的潜在优势至关重要。这样做可以提高图机器学习方法在大规模知识图谱上的效率和效果。这也带来了一个研究机会：探索适用于大型知识图谱的图机器学习方法的不同元采样方法。

B. 图机器学习模型与知识图谱的无缝集成

在 RDF 引擎之上实现图机器学习面临着重大挑战，主要在于训练后的模型与底层数据管理引擎之间的接口。常见的一种方法是使用用户定义函数（用户定义函数）来实现这一接口 [29]–[31]。然而，这会影响数据系统中的查询优化 [32]。大量用户定义函数的存在会限制机器学习相关查询的表达能力。例如，大量用户定义函数会让用户难以在这些函数中进行选择，并找到

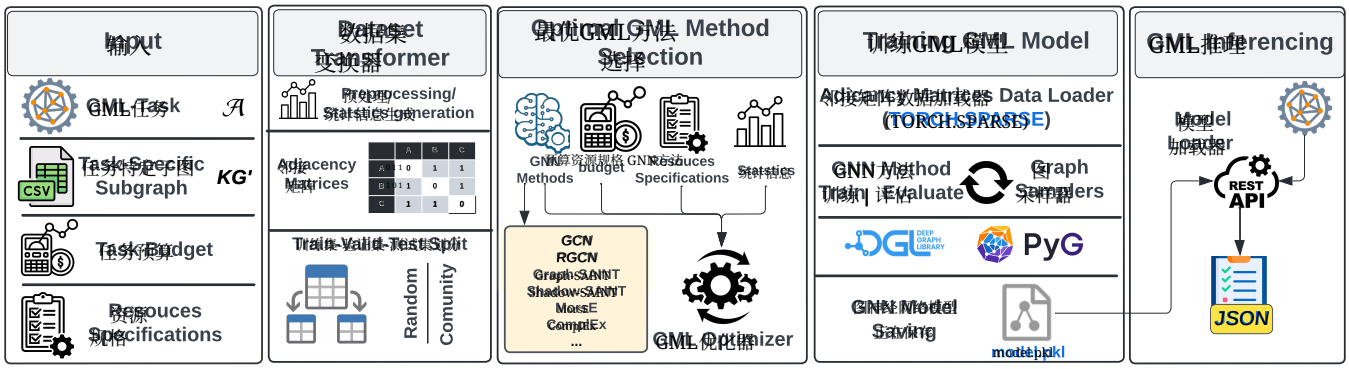


图6：我们提出的图机器学习即服务（GMLaaS）中训练流水线和推理的自动化。GMLaaS 与 KGMeta 管理器交互，在预算有限的情况下为特定任务训练模型。自动化流水线会选择近似最优的图机器学习方法，在预算有限的情况下训练模型。GMLaaS 支持通过由用户定义函数调用的 RestAPI 进行任务推理。

适合他们需求的那个。大多数现有查询优化器都没有用于估算这些用户定义函数开销的模型。因此，自动优化 SPARQL^{ML} 查询颇具挑战性。如何将训练后的图机器学习模型与资源描述框架无缝集成，仍有待研究。为应对这些挑战，我们提出了 KGMeta，将训练模型的元数据表示为图，并使其与知识图谱相互链接。

C. 优化 *SPARQL*^{ML} 查询和基准测试

用户定义谓词最初是为 SQL [33] 提出的。在 SPARQL^{ML} 中，用户定义谓词用于从与图中特定节点关联的某个训练模型获取预测。与估算传统资源描述框架谓词的开销相比，估算用户定义谓词的求值开销更为复杂。基数估算只用于优化后者的执行时间，而 SPARQL^{ML} 查询中的用户定义谓词则可以由多个模型进行推理，这些模型的准确率和推理时间各不相同。RDF 引擎无法获知这些信息，因此会面临如何选择最佳推理模型的问题。

对于 SPARQL^{ML} 查询，RDF 引擎使用选定模型执行推理步骤是一项颇具挑战的任务，需要进行优化，尤其是对推理过程进行排序。难点在于，需决定是在一次用户定义函数调用中执行推理，还是针对每个实例分别执行；后者可能导致大量用户定义函数调用。此外，每个模型都有独特的基数，即其能够生成的预测总数。因此，预测排序情况颇为复杂，因为 RDF 引擎无法准确估算用户定义函数的开销。

为应对这些挑战，研究人员可以开发基准测试，以评估 SPARQL^{ML} 查询的优化方法。这些基准测试应考虑针对不同用户定义谓词的各种模型，并能够适用于大型数据集。此外，每个 SPARQL^{ML} 查询中用户定义谓词的数量应有所不同，并且应关联基数不同的变量。这样便能全面评估 SPARQL^{ML} 查询中各种优化方法的性能和可扩展性。

IV. KGNET 平台

如图 3 所示，KGNet 提供两项主要服务，即图机器学习即服务（GMLaaS），以及基于现有资源描述框架引擎的 SPARQL^{ML} 即服务。

A. 图机器学习即服务（*GMLaaS*）

KGNet 是一个平台，可提供知识图谱上图机器学习训练的端到端自动化，如图 6 所示。该平台提供*GMLaaS*，这是一种 RESTful 服务，可通过自动训练和交互式推理来管理图机器学习模型。此外，它还利用嵌入存储，通过计算嵌入向量之间的相似度来支持实体相似性搜索任务。*GML* 训练管理器 会针对每项任务自动执行训练流水线。然而，由于知识图谱的复杂性和规模，在知识图谱上自动进行图机器学习训练颇具挑战。因此，KGNet 采用我们的元采样方法，通过选取适合给定任务的任务特定子图（KG'）来优化训练过程。此步骤有助于减少所需时间和内存，同时不牺牲准确率。该流水线的输入包括任务特定子图（KG'）、图机器学习任务、任务预算，以及机器学习环境中的可用资源。

数据转换器步骤会将子图转换为适用于内存中处理和矩阵运算的稀疏矩阵格式。该格式兼容常见的图机器学习数据加载器，例如 Py-Geometric 和 DGL，非常适合稀疏知识图谱。我们的流水线会验证节点/边类型的数量、移除字面量数据和目标类别边，并生成图统计信息，以确保数据一致性。我们还会采用随机划分和基于社区的划分等不同策略，将数据划分为训练集、验证集和测试集。KGNet 可自动完成这一转换，使临时图机器学习训练查询成为可能。

“最优图机器学习方法选择”步骤会为给定任务选择最佳的图机器学习方法。KGNet 支持多种图神经网络方法，包括 GCN、关系图卷积网络、Graph-SAINT、Shadow-SAINT、Morse，以及 ComplEx 等知识图谱嵌入方法。我们根据生成的稀疏矩阵的大小和数量，估算每种方法所需的内存；并根据矩阵维度和特征聚合方式，估算训练时间。此外，我们还估算训练时间

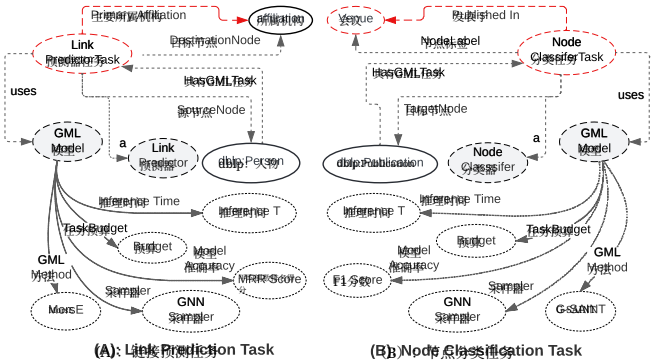


图 7：用于节点分类和链接预测任务的两个训练模型的 KGMeta 图。白色节点是来自原始 dataKG 的节点。虚线节点/边表示针对每个训练模型收集的元数据。

```
prefix dblp:<https://www.dblp.org/>
prefix kgnet:<https://www.kgnet.com/>
Insert into <kgnet>  { ?s ?p ?o }
where {select * from  kgnet.TrainGML(
{Name: 'MAG_Paper-Venue_Classifer',
 GML-Task:{ TaskType:kgnet:NodeClassifier,
 TargetNode: dblp:publication,
 NodeLable: dblp:venue},
 Task Budget:{ MaxMemory:50GB, MaxTime:1h,
               Priority:ModelScore} } )};
```

图 8：一条 SPARQL^{ML} 插入查询，用于在 DBLP 上训练论文-场所分类器。TrainGML 函数是一个在 RDF 引擎内部实现的用户定义函数。

具体而言，我们根据稀疏矩阵的维度，以及各方法采用的 GNN 邻居节点特征聚合方式进行估算。对于基于 GNN 采样的方法，采样成本基本取决于所使用的采样启发式方法 [34]。因此，我们正在研究一种更先进的估算方法：对稀疏矩阵进行采样，并在其上运行若干轮训练。

KGNet 的 GML优化器会为每种方法确定所需资源并优化训练设置，确保在分布式环境中的可扩展性。自动化流水线会训练模型，并收集评估指标和推理时间统计信息。系统会为训练模型生成一个 URI，以便将其与用于推理任务的其他模型区分开来。模型元数据会返回给 KGMeta 管理器，以更新 KGMeta 图。图7.a 和 b 分别展示了为链接预测模型和节点分类模型生成的元数据。图 3 所示的 *嵌入存储* 子组件通过存储、索引和搜索嵌入来实现快速相似性搜索。*GML推理* 接收用于推理的 HTTP调用，将结果序列化为 JSON RESTful API 响应，并将其发送回 RDF 引擎，如图 3 所示。当前版本使用 FAISS 嵌入存储 [35]，以支持节点相似性搜索的临时查询。

```
prefix dblp:<https://www.dblp.org/>
prefix kgnet:<https://www.kgnet.com/>
delete {?NodeClassifier ?p ?o}
where {
?NodeClassifier a kgnet:NodeClassifier.
?NodeClassifier kgnet:TargetNode dblp:Publication.
?NodeClassifier kgnet:NodeLabel dblp:venue.}
```

图 9：一条 SPARQL^{ML} 删除查询，用于删除训练模型及其元数据。

```
prefix dblp: <https://www.dblp.com/>
prefix kgnet: <https://www.kgnet.com/>
select ?author ?affiliation
where { ?author a dblp:person.
?author ?LinkPredictor ?affiliation.
?LinkPredictor a kgnet:LinkPredictor.
?LinkPredictor kgnet:SourceNode dblp:person.
?LinkPredictor kgnet:DestinationNode dblp:affiliation.
?LinkPredictor kgnet:TopK-Links 10.}
```

图 10：一条 SPARQL^{ML} 查询，用于预测 DBLP 知识图谱中的作者隶属关系链接（边）。

B. SPARQL^{ML} 即服务

我们提供 SPARQL^{ML} 即服务，由三个主要组件构成：查询管理器、KGMeta Governor 和元采样器。此外，我们还提供一种名为 SPARQL^{ML} 的接口语言，使用户能够使用类似 SPARQL 的查询来执行插入、删除或查询操作，具体如下：(*i*) SPARQL^{ML} 插入 查询用于训练图机器学习模型，并在 KGMeta 中维护其元数据（如图 8 所示）；(*ii*) SPARQL^{ML} 删除 查询用于从 GML-aaS 组件中删除训练模型文件及相关嵌入，然后从 KGMeta 中删除其元数据（如图 9 所示）；(*iii*) SPARQL^{ML} 查询查询用于查询知识图谱并进行推理，例如图 10 中的查询。收到 SPARQL^{ML} 查询后，查询管理器会对其进行解析。插入或删除查询将发送给 KGMeta Governor。如果是查询操作，则会先进行优化，再重写为 SPARQL 查询。

1) KGMeta Governor：KGMeta Governor 会为每个知识图谱维护一个 KGMeta 图，并使用该知识图谱专属的已训练图机器学习模型所收集的统计信息和元数据。INSERT 查询是针对某个知识图谱训练任务的请求。解析所得的信息包括任务类型（例如节点分类或链接预测）、任务输入（例如分类任务中的目标节点和分类标签（Y 类）），以及预算（例如内存和时间预算）。经验丰富的机器学习用户还可以提供其他信息，例如超参数或特定的图机器学习方法。这些信息会封装为一个 JSON 对象，如图 8 所示。在第 4 行，*TrainGML* 是一个用户定义函数，其输入为封装了训练图机器学习模型所需全部信息的 JSON 对象。KGMeta Governor 会将任务发送给元采样器，以获取适用于给定任务的任务特定子图（KG'）。随后，Governor 会与 GML 训练管理器交互，以自动执行训练

```
prefix dblp: <https://www.dblp.org/>
prefix kgnet: <https://www.kgnet.com/>
select ?title
       sql:UDFS.getNodeClass($m,?paper) as ?venue
where {
    ?paper a dblp:Publication.
    ?paper dblp:title ?title.
}
```

图 11：用于 SPARQL$_{pv}^{\mathbf{ML}}$ 的候选 SPARQL

该任务的流水线。训练完成后，KGMeta Governor 会接收训练模型的元数据，包括准确率和推理时间，以维护 KGMeta，如图7所示。

2）元采样器：我们的元采样器旨在为训练图机器学习任务识别任务特定子图（KG'）。每个图机器学习任务的目标是特定类型的节点，例如 SPARQL$_{pv}^{\mathbf{ML}}$ 中的 dblp:Publication。我们的元采样器会提取任务特定子图（KG'），其中包含一组三元组，以及与目标节点相关的代表性三元组。根据知识图谱模式结构，KG'的大小远小于知识图谱的大小。较小的规模可缩短图机器学习任务 $\mathcal{A}$的训练时间，并减少训练所需的内存。然而，知识图谱中可能包含无法从目标节点 v^T 到达，或需经过超过三跳才能与 v^T相连的三元组。这些三元组无助于模型的泛化，还可能导致过平滑问题 [36], [37]。

我们的基于SPARQL的元采样方法通过两个参数确定所提取子图的范围：(*i*) 方向 d，其中 $d=1$ 表示出边方向，$d=2$ 表示双向（即同时包括出边和入边）；以及 (*ii*) 跳数 h。我们使用 $d \in \{1,2\}$ 和 $h \in \{1,2\}$ 的四种组合评估了该方法的性能。对于节点分类，我们的元采样方法在 $d=1$ 和 $h=1$ 时取得了更好的结果；对于链接预测，则在 $d=2$ 和 $h=1$ 时表现更佳。

3) 查询管理器：查询管理器负责优化用于模型选择和排序的 SPARQL$^{\mathbf{ML}}$ 查询，以评估用户定义谓词。对于图 2 所示的 SPARQL$_{pv}^{\mathbf{ML}}$，查询优化器会获取满足与用户定义谓词 *?NodeClassifier* 相关条件的所有模型 URI。KGMeta 是一个资源描述框架图，其中包含优化器统计信息，例如模型准确率、推理时间和模型基数。因此，我们使用 SPARQL 查询获取模型的 URI、准确率、推理时间和基数。查询优化器会选择准确率高且推理时间短的近似最优图机器学习模型。我们将此问题定义为一个整数规划优化问题，以最小化总执行时间或最大化推理准确率。

SPARQL$^{\mathbf{ML}}$ 查询重写器使用带有 URI m 的近似最优图机器学习模型，生成候选 SPARQL 查询。KGNet 目前支持两种可能的执行计划，其查询模板如图11和图12所示。核心思路是将用户定义谓词映射为用户定义函数（UDF），例如 *sql:UDFS.getNodeClass*，以便在执行期间向图机器学习推理发送 HTTP调用

```
prefix dblp: <https://www.dblp.org/>
prefix kgnet: <https://www.kgnet.com/>
select ?title
  sql:UDFS.getKeyValue(?venues_dic,?paper) as ?venue
where {
    ?paper a dblp:Publication.
    ?paper dblp:title ?title.
{select sql:UDFS.getNodeClass($m,dblp:Publication)
  as ?venues_dic where { } }}
```

图 12：SPARQL$_{pv}^{\mathbf{ML}}$的候选 SPARQL

在我们的 GMLaaS 中，Manager 会基于预训练模型 m进行推理。HTTP调用的数量可能会主导查询执行成本。例如，SPARQL$_{pv}^{\mathbf{ML}}$会预测所有论文的会议地点，其规模为 |?papers|。

图 11 所示的查询模板将生成|?papers| 次 HTTP调用。然而，图 12 所示的查询模板通过强制执行一个内部选择查询来构建包含所有论文及其预测会议地点的字典，将 HTTP调用次数减少到一次。然后，使用 *sql:UDFS.getKeyValue* 查找每篇论文的会议地点。我们的查询优化器会将 SPARQL$^{\mathbf{ML}}$ 查询中与查询知识图谱三元组相关的三元组模式，按与用户定义谓词关联的变量划分为多个集合。例如，在图 2 所示的 SPARQL$_{pv}^{\mathbf{ML}}$查询中，我们的优化器会识别出两个与变量 *?paper* 匹配的三元组模式，以及一个与变量 *?venue*匹配的三元组模式。我们使用 SPARQL 查询来获取每个集合的基数，即数据集中该变量不同取值的数量。我们将此问题表述为另一个整数规划优化问题 [38]，以最小化 HTTP调用总数或构建的字典大小，而字典大小取决于模型基数。例如，对于图 12 所示的查询，我们的优化器会生成一个包含所有论文及其预测会议地点的字典，然后使用用户定义函数*sql:UDFS.getKeyValue*来检索每篇论文的会议地点。

五、实验评估

本节分析 KGNet 自动化流水线的能力：与针对完整图的传统流水线相比，它能以更少的时间和内存，为特定任务训练模型。

A. 评估设置

对比方法：我们将 RGCN [23] 用作全批次训练方法，将 GraphSAINT [20],和 ShadowSAINT [21]用作基于小批量采样的节点分类方法，并将 MorsE [22] 用作基于边采样的链接预测方法。采样和训练均采用 OGB [19] 默认配置。所有实验中的节点特征均通过 Xavier 权重初始化随机生成。

计算基础设施：所有实验均在 Ubuntu 服务器虚拟机上进行，该虚拟机配备两颗 64 核英特尔至强 2.4 GHz（Skylake，IBRS）CPU、256 GB 主内存和 1 TB 磁盘存储。

真实知识图谱：我们主要关注两个基准知识图谱，它们在图规模、图数据领域、任务类型和连接密度方面各不相同，包括（DBLP [11] 和 Yago-4

表 I：所用知识图谱和图神经网络任务的统计信息。我们使用的知识图谱（DBLP 和 Yago）规模是 OGB [19] 中所报告数据集的四倍。

知识图谱	DBLP	YAGO4
#三元组	252M	400M
#目标	50 个会议 51K 个隶属机构 1.2M 篇论文	200 个国家
#边类型	48	98
#节点类型	42	104
任务	NC,LP,ES	NC

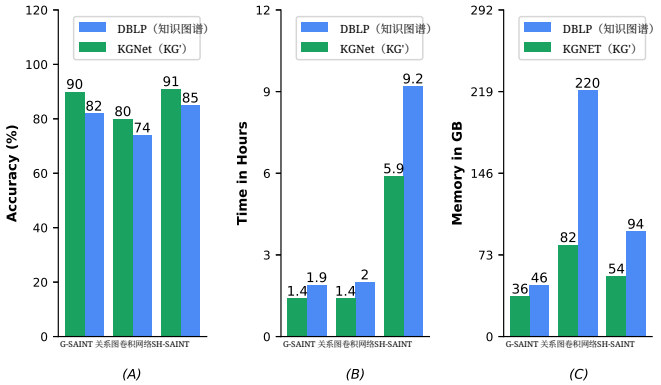


图 13：(a) DBLP 知识图谱论文-会议节点分类任务的准确率、(B) 训练时间和 (C) 训练内存。KGNet 面向任务采样的子图 (KG') 显著提升了准确率、缩短了训练时间并减少了内存占用。

[39]）。我们开展了两个节点分类任务和一个链接预测任务。我们采用了 OGB [19]中使用的任务。所用知识图谱和任务的统计信息见表 I。

端点：我们使用 Virtuoso 07.20.3229 作为 SPARQL 端点，因为它被广泛用作大型知识图谱的端点，例如 DBLP。所有实验均在端点上运行 Virtuoso 引擎的标准、未经修改的安装版本。

B. 使用真实知识图谱进行图机器学习实验

我们开展了三项图机器学习任务，以评估 KGNet 自动化图机器学习流水线。对于**节点分类任务**，我们使用图神经网络方法训练节点分类器，为每篇 DBLP 论文预测其所属会议。知识图谱被加载到 Virtuoso 资源描述框架引擎中。KGNet 使用 $d1h1$ 执行元采样，以提取任务特定子图（KG'）用于训练关系图卷积网络、Graph-SAINT 和 Shadow-SAINT 方法。图 13 和图 14 中的任务结果表明，在所有方法中，我们使用（KG'）的 KGNet 训练流水线均优于基于完整知识图谱的传统流水线，准确率最高提升了 11%。KGNet 的自动化训练流水线成功帮助图神经网络方法显著降低了内存消耗和训练时间。具体而言，KGNet 的内存消耗至少降低了 22%，训练时间至少缩短了 27%。这些结果表明，KGNet 能够为每项任务有效发现任务特定子图。

我们的**链接预测任务**旨在根据作者的出版物和所属机构历史记录，预测其在

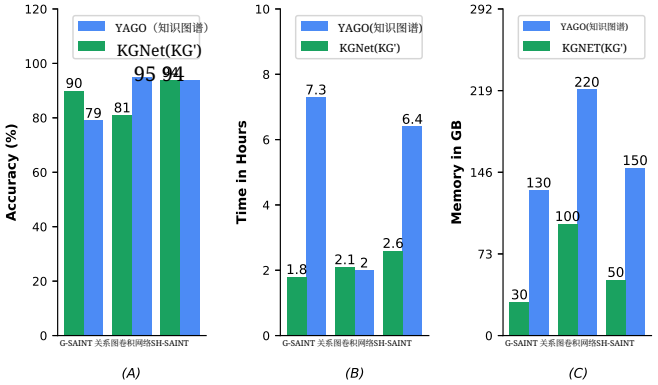


图 14：YAGO-4 知识图谱地点-国家节点分类任务的 (a) 准确率、(B) 训练时间和 (C) 训练内存。KGNet 面向任务采样的子图（KG'）显著提升了准确率，并缩短了训练时间、降低了内存占用。

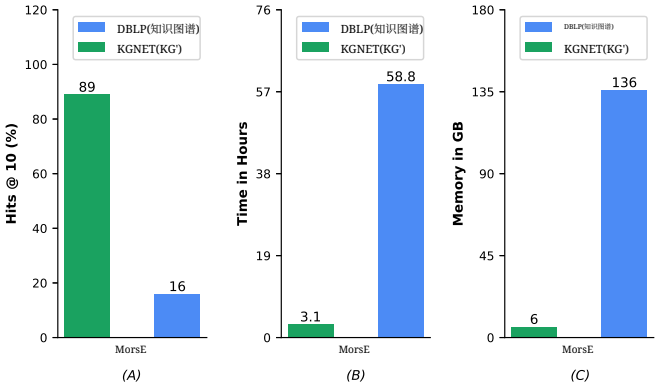


图 15：DBLP 作者隶属关系链接预测任务的 (a) 准确率、(B) 训练时间和 (C) 训练内存。KGNet 面向任务的边采样子图（KG'）显著提升了 Hits@10 MRR 分数，并缩短了训练时间、降低了训练内存占用。

DBLP知识图谱上的所属机构链接。MorsE [22]是最先进的基于采样的链接预测方法。我们在传统流水线中使用完整知识图谱运行 MorsE。在 KGNet 流水线中，我们的元采样首先使用 $d2h1$ 提取任务特定子图（KG'），以训练 MorsE。如图 15 所示，结果表明，在 Hits@10 MRR 分数,方面，KGNet 自动化流水线优于基于完整知识图谱训练的流水线。KGNet 大幅降低了内存使用量和训练时间，与基于完整知识图谱训练的流水线相比，降幅达 94%。

VI. 相关工作

人工智能与数据库系统相结合的应用正迅速增长，主要形成了两种方法：将人工智能模型集成到数据库系统中（AI4DB），以及优化数据库技术以服务人工智能系统，从而提升可扩展性（DB4AI） [40]。在 KGNet 中，我们将 SPARQL^{ML} 归为 AI4DB 方法，因为我们扩展了知识图谱引擎，使其能够使用图机器学习查询知识图谱并进行推理

模型。不过，我们将 GMLaaS 归为 DB4AI 方法，因为我们采用元采样方法优化了训练流水线，该方法通过查询知识图谱来提取特定任务的子图。WorksRDFFrames [41], DistRDF2ML[42],和 Apple Saga [18] 旨在弥合机器学习与 RDF 系统之间的差距，使用户能够以标准表格格式从异构图引擎中提取数据，以执行分类、回归和聚类等传统机器学习任务，或使用知识图谱嵌入方法生成节点/边嵌入，用于相似性搜索应用。

Yuo Lu 等人在 [29] 中研究了 SQL 的人工智能赋能查询优化问题，并提出了概率谓词（PP）方法，该方法无需了解推理模型即可进行训练。在 Learned B+树 [43], 中，B+树索引基于人工智能模型进行优化，这些模型将每个查询键映射到对应的页。Hasan 等人 [44] 利用自回归密度模型表示各列之间的联合数据分布，从而实现快速连接查询。ITLCS [45]提出了一种基于机器学习的索引选择方法，该方法结合分类器模型和遗传算法，以选出准确的索引。Stardog [13] 支持监督学习，可用于构建预测分析模型。Stardog 允许用户编写 SPARQL 查询，以表格格式收集机器学习训练特征集，并应用经典机器学习方法，即分类、聚类和回归，以支持推理查询。

Google 的 BigQuery ML [46]通过引入一种混合语言模型，为 SQL 语句中的人工智能模型提供易用工具。该模型同时包含人工智能操作和数据库操作，并在 TensorFlow、Keras 等人工智能平台上执行人工智能操作。SQL4ML [47] 将以 SQL 实现的机器学习运算符转换为 TensorFlow 流程，以实现高效训练。为了通过 SPARQL 支持临时图机器学习流水线，资源描述框架引擎需要具备这一级别的支持。

Bordawekar 等人 [48]构建了一种认知关系数据库引擎，利用基于 word2vec 嵌入的词语相似性查询数据库记录，并通过外部数据源扩展查询结果。认知数据库代表着将表示学习与数据库相结合的一步，其采用了文本嵌入技术。EmbDI [49] 利用词嵌入技术，从关系数据集中自动学习高质量的局部关系嵌入，以支持数据集模式匹配。与上述所有系统不同，KGNet 提出了一种结合 DB4AI 和 AI4DB 方法的平台，以弥合图机器学习框架与资源描述框架引擎之间的差距。

七、结论

图机器学习框架与 RDF 引擎缺乏集成，因此数据科学家必须手动优化图机器学习流水线，以检索存储在 RDF 引擎中的知识图谱，并选择与其计算资源相匹配的图机器学习方法。此外，训练后的模型无法直接用于知识图谱上的查询和推理，这阻碍了系统的可扩展性，尤其是在知识图谱规模不断增大、需要大量计算资源时。此外，这些限制也影响了系统的灵活性，因为描述性查询语言无法纳入图机器学习模型。为克服这些限制，这篇愿景论文提出了 KGNet 是一个构建于 RDF 引擎之上的按需图机器学习即服务（GMLaaS）平台，旨在支持图机器学习增强的 SPARQL 查询（SPARQL^{ML}）。KGNet 使用元采样提取任务特定子图（KG'），将其作为针对知识图谱中特定任务的搜索查询。我们的图机器学习即服务可自动构建经济高效的流水线，使用 KG' 在给定的时间或内存预算内训练模型。KGNet 将训练后模型的元数据和统计信息维护为名为 KGMeta 的 RDF 图，并将其与关联知识图谱一同存储。KGMeta 实现了图机器学习模型与 RDF 引擎之间的无缝集成，使用户能够基于模式匹配的 SPARQL 逻辑轻松编写 SPARQL^{ML} 查询。此外，KGMeta 还能让 KGNet 优化 SPARQL^{ML} 查询，以便在推理过程中进行模型选择和排序。KGNet 为数据管理和人工智能领域带来了新的研究机遇。

参考文献

[1] H. Aidan、B. Eva 等，《知识图谱》，ACM Comput. Surv.，第54卷，第4期，2021年。 [在线]。可访问：https://doi.org/10.1145/ 3447772[2] S. Wu、F. Sun、W. Zhang、X. Xie 和 B. Cui，《推荐系统中的图神经网络：综述》，ACM Comput. Surv.，第55卷，第5期，第97:1–97:37页，2023年。 [在线]。可访问：https://doi.org/ 10 1145 3535101 。/ Z. Wang、Q. Lv、X. Lan 和 Y. Zhang，[3]《通过图卷积网络实现跨语言知识图谱对齐》，载于自然语言处理实证方法会议（EMNLP）论文集，2018年，第349–357页。 [在线]。可访问：https://aclanthology.org/D 18 1032 - X. Lin、Z. Quan、Z. Wang、T. Ma 和 X. Zeng，[4]载于国际人工智能联合会议（IJCAI）论文集，2020年，第2739–2745页。 [在线]。可访问：https://doi.org/10.24963/ijcai.2020/ 380[5] Y. Dou、Z. Liu、L. Sun、Y. Deng、H. Peng 和 P. S. Yu，《增强基于图神经网络的欺诈检测器，以抵御伪装欺诈者》，载于ACM国际信息与知识管理会议（CIKM），2020年，第315–324页。 [在线]。可访问：https://doi.org/ 10 1145 3340531 3411903 。/ . Y. Dou、Z. Liu 等，[6]《增强基于图神经网络的欺诈检测器，以抵御伪装欺诈者》。ACM，2020年。 [在线]。可访问：https://doi.org/ 10 1145 3340531 3411903 。/ . PyG。（）Torch [7] 2022 [Geometric 文档。在线]。可访问：https://pytorch-geometric.readthedocs.io/en/latest/index.html [8] D. Zheng、C. Ma、M. Wang 等，《DistDGL：面向十亿级图的分布式图神经网络训练》，载于 IEEE/ACM，第10届。IEEE，2020年。[9] Z. Wu、S. Pan、F. Chen 等，《图神经网络综合综述》，IEEE《神经网络与学习系统汇刊》，第32卷，第1期，第4–24页，2021年。 [在线]。可访问：https: //doi.org/ 10 1109 2020 2978386 。/TNNLS。 . Q. Wang、Z. Mao、B. Wang 和 L. Guo，[10]《知识图谱嵌入：方法与应用综述》，IEEE《知识与数据工程汇刊》，第29卷，第12期，第2724–2743页，2017年。[11] M. R. Ackermann。（2022年）DBLP 中的资源描述框架（RDF）。 [在线]。可访问：https: //blog.dblp.org/2022/03/02/dblp-in-rdf/ [12] D. Gordon。（2021年）Neo4j 中的全文搜索。 [在线]。可访问：https://neo4j.com/developer/kb/fulltext-search-in-neo4j/ [13] docs@stardog.com。Stardog 文档——机器学习。 [在线]。可访问：https://docs.stardog.com/machine-learning#machine-learning [14] Neo4j.Inc。（2022年）面向机器学习的 Neo4j 图数据科学。[在线]。可访问：https://neo4j.com/docs/graph-data-science/current/machine-learning/machine-learning/ [15] H. Chen、S. F. Sultan 等，《通过极稀疏随机投影实现快速准确的网络嵌入》，载于第28届ACM国际信息与知识管理会议论文集，系列号 CIKM '19，2019年，第399–408页。 [在线]。可访问：https://doi.org/10.1145/3357384.3357879

[16] A. Grover 和 J. Leskovec，《node2vec：面向网络的可扩展特征学习》，载于第22届ACM SIGKDD，2016年，第 855 864 [ ] 10 1145 2939672 2939754 – 页。在线。[17] 可访问：https://doi.org/。/。W. L. Hamilton、Z. Ying 和 J. Leskovec，《大型图上的归纳式表示学习》，载于 NeurIPS，2017年，第1024–1034页。 [在线]。可访问：https://proceedings.neurips.cc/paper/2017/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html[18] I. F. Ilyas、T. Rekatsinas、V. Konda、J. Pound、X. Qi 和 M. A. Soliman，《Saga：用于大规模持续构建和提供知识的平台》，载于 SIGMOD，2022年，第2259–2272页。 [在线]。可访问：https://doi.org/ 10 1145 3514221 3526049 。/。[19] W. Hu、M. Fey、M. Zitnik、Y. Dong、H. Ren、B. Liu、M. Catasta 和 J. Leskovec，《开放图基准：用于图机器学习的数据集》，载于 NeurIPS，2020年。[20] H. Zeng、H. Zhou 等，《GraphSAINT：一种基于图采样的归纳学习方法》，载于 ICLR(8)，2020年。 [在线]。可访问：https://openreview.net/forum?id=BJe8pkHFwS [21] H. Zeng、M. Zhang、Y. Xia 等，《解耦图神经网络的深度与范围》，CoRR，第abs/2201.07858卷，2022年。 [ ] 2201 07858 在线。可访问：[22] https://arxiv.org/abs/。M. Chen、W. Zhang 等，《面向归纳式知识图谱嵌入的元知识迁移》，载于 ACM SIGIR，2022年，第927–937页。 [ ] 10 1145 3477495 3531757 在线。可访问：[23] https://doi.org/。/。M. S. Schlichtkrull、T. N. Kipf 和 Peter Bloem 等，《使用图卷积网络对关系数据建模》，载于 ESWC，第10843卷。Springer，2018年，第593–607页。 [在线]。可访问：https://doi.org/ 10 1007 978 3 319 93417 4_38 。/ - - - - M. Chen、Y. Zhang、X. [24] Kou 等，《R-GAT：面向多关系图的关系图注意力网络》，CoRR，第abs/2109.05922卷，2021年。 [ ] 2109 05922 在线。可访问：https://arxiv.org/abs/。M. Färber，《微软学术知识图谱：包含80亿学术数据三元组的关联数据源》，载于 ISWC，第11779卷，2019年，第113–129页。 [在线]。可访问：https://doi.org/10.1007/ 978-3-030-30796-7 8 [25] [26] T. Y. Cheng、Q. Hu、Q. Xie、N. Trigoni 和 A. Markham，《元采样器：近乎通用且面向任务的点云采样》，ECCV，2022年。 [在线]。可访问：https://doi.org/10.48550/arXiv.2203. 16001[27] O. Dovrat、I. Lang 和 S. Avidan，《学习采样》，载于 IEEE CVPR，2019年，第2760–2769页。[在线]。可访- 问：http://openaccess.thecvf.com/content_CVPR_2019/html/Dovrat_Learning_to_Sample_CVPR_2019_paper.html [28] Y. Xiao、K. Gong、P. Zhou、G. Zheng、X. Liang 和 L. Lin，《面向多语言低资源语音识别的对抗式元采样》，载于 AAAI，2021年，第14 112–14 120页。 [在线]。可访问：https://ojs.aaai.org/index.php/AAAI/article/view/ 17661 [29] Y. Lu、A. Chowdhery、S. Kandula 和 S. Chaudhuri，《利用概率谓词加速机器学习推理》，载于 SIGMOD，系列号 SIGMOD '18，2018年，第1493–1508页。 [在线]。可访问：https: // doi.org/ 10 1145 3183713 3183751 。/。R. Bordawekar 和 O. Sh mueli，[30] 《在关系数据库中利用低维词嵌入实现认知智能查询》，CoRR，第abs/1603.07185卷，2016年。 [在线]。可访问：http://arxiv.org/ abs/1603.07185 [31] M. Schule、H. Lang、M. Springer、A. Kemper、T. Neumann 和 S. Gunnemann，《使用 SQL 和 GPU 在数据库内进行机器学习》，载于 SSDBM，2021年。 [在线]。可访问：https://doi.org/10.1145/3468791. 3468840[32] K. Awada、M. Y. Eltabakh、C. Tang、M. Al-Kateb、S. Nair 和 G. Au，“Teradata IntelliSphere 中异构 SQL 大数据基础设施的成本估算”，载于 EDBT，2020年。[33] S. Chaudhuri 和 K. Shim，“使用用户定义谓词优化查询”，ACM Trans. Database Syst.，第 24， 2， 177 228, 卷，第期，页码 – 。在线。获取地址：1999 [ ] 10 1145 320248 320249 。/ 。[34] M. Serafini，“可扩展的图神经网络训练：采样的价值”，ACM SIGOPS Oper. Syst. Rev.，第 55， 1， 68 76, 卷，第期，页码 – 。在线。获取地址：https://doi.org/ 。/ 。2021 [ ] 10 1145 3469379 3469387 [35] J. Johnson、M. Douze 和 H. Jégou，“使用 GPU 进行十亿规模的相似性搜索”，IEEE Trans. Big Data，第 7 卷，第 3 期，页码 535–547，2021年。 [ ] 10 1109 2019 2921572 在线。获取地址：[36] https://doi.org/ 。/TBDATA. 。D. Chen、Y. Lin、W. Li、P. Li、J. Zhou 和 X. Sun，“从拓扑视角衡量并缓解图神经网络的过平滑问题”，载于 IAAI，2020年，页码 3438–3445。[37] J. Liu、K. Kawaguchi、B. Hooi、Y. Wang 和 X. Xiao，“Eignn：高效的无限深度图神经网络”，载于 NIPS，第 34 卷，2021年，页码 18 762–18 773。 [在线]。获取地址：https://proceedings.neurips.cc/paper/2021/file/9bd5ee6fe55aaeb673025dbcb8f939c1-Paper.pdf [38] S. Bradley、A. Hax、A. Hax 和 T. Magnanti，第 9 章：整数规划。Addison-Wesley，1977年。[39] T. P. Tanon、G. Weikum 和 F. M. Suchanek，“YAGO 4：一个合理的知识库”，载于 ESWC，第 12123 卷，2020年，页码 583–596。 [在线]。获取地址：https://doi.org/ 10 1007 978 3 030 49461 2_34 。/ - - - - X. Zhou、C. Chai、[40] G. Li 和 J. Sun，“数据库与人工智能相遇：综述”，IEEE TKDE，第 34 卷，1116，2022年 3，1096 第期，页码 – 。[41] A. Mohamed、G. Abuoda 等，“Rdfframes：面向机器学习工具的知识图谱访问”，PVLDB，第 13 卷，第 12 期，2020年。[在线]。获取地址：http://www.vldb.org/pvldb/vol13/p2889-mohamed. pdf [42] C. F. Draschner、C. Stadler、F. Bakhshandegan Moghaddam、J. Lehmann 和 H. Jabeen，“DistRDF2ML：面向 RDF 知识图谱的可扩展分布式内存机器学习流水线”，载于第 30 届 ACM 会议论文集，2021年，页码 4465–4474。 [在线]。获取地址：https: //doi.org/ 10 1145 3459637 3481999 。/ 。T. Kraska、A. Beutel、E. H. Chi、J. Dean 和 [43] N. Polyzotis，“ 学习型索引结构的价值 ”，载于 SIGMOD。ACM，2018 年，页码 489–504。 [ ] 10 1145 3183713 3196909 在线。获取地址：[44] https://doi.org/ 。/ 。S. Hasan、S. Thirumuruganathan、J. Augustine、N. Koudas 和 G. Das，“使用深度学习进行多属性选择率估计”，CoRR，第 abs/1903.09999 卷，2019年。[在线]。获取地址：http://arxiv.org/abs/1903. 09999[45] W. G. Pedrozo、J. C. Nievola 和 D. C. Ribeiro，“一种利用学习分类器系统在混合存储环境中进行索引调优的自适应方法”，载于 HAIS，第 10870 卷，2018年，页码 716–729。 [在线]。获取地址：https://doi.org/ 10 1007 978 3 319 92639 1_60 。/ - - - - S. Fernandes 和 J. [46] Bernardino，“什么是 BigQuery？ ”，《第 19 届国际数据库工程与应用研讨会论文集》，2015年。[47] N. Makrynioti、R. Ley-Wild 和 V. Vassalos，“SQL4ML：面向机器学习的声明式端到端工作流”，CoRR，第 abs/ 1907 12415，卷，。在线。获取地址：2019 [ ] 1907 12415 http://arxiv.org/abs/ 。[48] R. Bordawekar、B. Bandyopadhyay 和 O. Shmueli，“认知数据库：赋予关系数据库人工智能能力的探索”，第 abs/1712.07199 卷，2017年。 [在线]。获取地址：http://arxiv.org/abs/ 1712 07199 。C. Riccardo、P. [49] Paolo 和 T. Saravanan，“为数据集成任务创建异构关系数据集的嵌入”，载于 ACM SIGMOD，美国，2020年，页码 1335–1349。 [在线]。获取地址：https://doi.org/10.1145/3318464.3389742