

群体知识图谱：分布式知识迁移与联邦式图谱推理

陈名杨¹, 张文², 陈湘楠², 周虹廷¹, 陈华钧¹

(1. 浙江大学计算机科学与技术学院, 浙江 杭州 310007;

2. 浙江大学软件学院, 浙江 杭州 310007)

摘要: 群体知识图谱是指通过群体协作, 以去中心化或分布式方式管理和维护的知识图谱。相比现有的集中式管理的知识图谱, 群体知识图谱具备知识确权、隐私保护、众包激励、可信溯源等特点。尝试探讨构建或应用群体知识图谱平台面临的技术挑战。其中分布式知识迁移考虑在一个分散自治的框架下, 通过实现不同来源的多个知识图谱之间的知识迁移, 缓解单个知识图谱的知识不完备问题。其主要难点是在充分保护知识的自治所有权的前提下, 尽可能共享有用的知识, 以增强各自的知识图谱表示。联邦式图谱推理也是考虑在一个分布式环境下, 通过联邦学习机制实现隐私保护前提下的知识图谱推理。在分布式知识迁移中, 强调在关系集合互相重叠的知识图谱间迁移与实体无关的知识; 而在联邦式图谱推理中, 强调在多个实体集合互相重叠的知识图谱间共同学习更好的实体嵌入表示。针对这两个问题分别进行模型设计及实验验证。

关键词: 群体知识图谱; 联邦学习; 知识迁移

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202217

Collective knowledge graph: meta knowledge transfer and federated graph reasoning

CHEN Mingyang¹, ZHANG Wen², CHEN Xiangnan², ZHOU Hongting¹, CHEN Huajun¹

1. College of Computer Science and Technology, Zhejiang University, Hangzhou 310007, China

2. School of Software Technology, Zhejiang University, Hangzhou 310007, China

Abstract: Collective knowledge graphs refer to knowledge graphs that are managed and maintained in a decentralized or distributed manner through group collaboration. Compared with the existing centrally managed knowledge graph, the collective knowledge graph has the characteristics of knowledge right confirmation, privacy protection, crowd sourcing incentive, and credible traceability. Tring to explore the technical challenges faced by building and applying a collective knowledge graph platform. For meta knowledge transfer, the knowledge incompleteness of a single knowledge graph by knowledge transfer among multiple knowledge graphs from different sources under a decentralized and autonomous framework was considered. The main difficulty was to enhance the respective knowledge graph representation by sharing useful knowledge with each other as much as possible while fully protecting the autonomous ownership of knowledge. For federated graph reasoning, the knowledge graph reasoning in a distributed environment under the privacy-preserving by means of the federated learning mechanism was considered. Meta knowledge transfer focused on transferring entity-independent knowledge between knowledge graphs with overlapped relation set, while federated graph reasoning aimed at learning better entity embeddings for knowledge graphs with overlapped entity set. The model design and experimental validation for each of these two problems were conducted.

Key words: collective knowledge graph, federated learning, knowledge transfer

收稿日期: 2021-12-27; 修回日期: 2022-02-25

通信作者: 陈华钧, huajunsir@zju.edu.cn

基金项目: 国家自然科学基金资助项目 (No.91846204, No.U19B2027)

Foundation Items: The National Natural Science Foundation of China (No.91846204, No.U19B2027)

0 引言

知识图谱(knowledge graph, KG)旨在用图的形式描述客观世界的实体、概念、事件及其之间的关系。实体是客观世界中的具体人物或事物,如篮球运动员姚明、浙江大学等。概念是指人对世界万物的主观概念抽象,如运动员、高等学校等。事件是指客观事物发生的行为及活动,如股价波动、机构兼并等。关系描述实体、概念、事件之间存在的关联关系,如张三毕业于浙江大学、大学是高等学校等。

知识图谱的发展历史源远流长,从经典人工智能阶段的知识工程,到互联网时代的语义 Web^[1],再到当下很多领域构建的数千亿级别的现代知识图谱^[2-4]。知识图谱在语义搜索^[5]、智能问答^[6]、推荐计算^[7]、语言理解^[8]、大数据分析^[9]、设备物联^[10]等领域都有广泛应用。

尽管知识图谱技术取得了飞速进步,但其仍面临众多挑战。知识图谱强调将分散的碎片化数据相互关联、链接和融合。即使在单一的企业或机构内部,知识图谱的数据也通常来源于不同部门,知识的生产者应该拥有知识的所有权和控制权,知识图谱的构建及应用通常需要依靠多部门群体协作完成。因此,一个好的知识图谱平台应该是分布式的,而非集中或中心化的。

群体知识图谱是指通过群体协作,以去中心化和分布式方式管理和维护的知识图谱。相比现有的集中式管理的知识图谱,群体知识图谱具备知识确权、隐私保护、众包激励、可信溯源等特点。知识确权指保护知识的所有权和控制权,这也是有效进行知识众包激励的保障。隐私保护指在知识被进一步应用和消费的过程中,更加有效地保护知识中的敏感信息。同时,由于错误、虚假甚至恶意的知识会给下游的知识使用者带来不利后果,因此还需要提供必要的机制来追溯知识的来源,即知识的可信与溯源^[10]。具体来说,群体知识图谱是多个在组成部分上相互关联的知识图谱(如部分相同的关系集合或实体集合),并且这些知识图谱可以进行去中心化和分布式管理和维护。

一般来说,传统的知识图谱具有不完整性,在构建过程中对知识图谱进行补全是一个重要的任务。同样地,在构建一个完整的群体知识图谱平台时,也面临众多新的技术挑战。本文尝试探讨其中的两个问题,即分布式知识迁移和联邦式图谱推

理。分布式知识迁移考虑在一个分散自治的框架下,通过实现不同来源的多个知识图谱之间的知识迁移,解决单个知识图谱的知识不完备问题。其主要难点是在充分保护知识的自治所有权的前提下,尽可能共享有用的知识,以增强各自的表示学习。联邦式图谱推理考虑在一个分布式环境下,通过联邦学习机制实现隐私保护前提下的知识图谱推理。其中分布式知识迁移强调将与实体无关的知识从源知识图谱转移到目标知识图谱,并且目标知识图谱中的关系集合为源知识图谱关系集合的子集。联邦式图谱推理强调利用联邦学习让多个知识图谱间共享的实体学习到更好的表示。前者多个图谱间具有重叠的关系集合,后者多个图谱间具有重叠的实体集合。本文将分别介绍这两个问题,并进行模型设计和实验验证。

本文主要有以下贡献。

(1) 强调群体知识图谱的重要性,并描述其相关特征。在其基础上提出分布式知识迁移和联邦式图谱推理两个问题。

(2) 分别设计模型,利用元学习解决分布式知识迁移,利用联邦学习解决联邦式图谱推理。

(3) 实验证明,设计的模型在上述两个问题场景下都能取得较好的结果,证明了模型设计的有效性。

1 相关工作

1.1 知识图谱表示学习

知识图谱表示学习方法通过学习知识图谱中实体和关系的向量表示(即知识图谱表示、知识图谱嵌入),对知识图谱中的固有语义信息进行向量化表示。这种向量化表示可以弥补以结构化三元组表示知识图谱的不足,并且可以被更方便地部署到许多与知识图谱相关的任务上。大多数知识图谱表示学习的工作都基于一些特定的关系模式假设,从而设计具有不同表达能力的打分函数,例如,基于翻译距离假设的模型^[11-13]和基于语义匹配假设的模型^[14-16]。当前的知识图谱表示学习方法大多针对单个知识图谱的场景,即考虑在一个知识图谱上学习该知识图谱中实体和关系的表示,并没有考虑在多个知识图谱的场景(如本文提出的群体知识图谱)下,学习实体和关系的表示。

1.2 分布式知识迁移

分布式知识迁移考虑在多个具有相同关系集合的知识图谱之间进行知识迁移,之前的方法可以

分为基于子图推理的方法和基于规则学习的方法。基于子图推理的方法包括图归纳学习（graph inductive learning, GraIL）^[17]、用于可归纳关系推理的交互信息传递网络（communicative message passing neural network for inductive relation reasoning, CoMPILE）^[18]和拓扑可知的关联模型（topology-aware correlation, TACT）^[19]，这类方法通过子图提取图神经网络，从而学习到独立于具体实体的关系预测能力，并且这种能力可以被推广到新知识图谱中未出现过的实体，从而完成知识迁移。另外，由于逻辑规则可以独立于具体的实体，因此基于规则学习的方法也可以处理新出现的实体之间的关系预测，从而完成知识迁移。基于规则学习的方法可以分为显式的规则学习以及端到端可微分的规则学习，前者包括在不完整信息下的联合规则挖掘（associate rule mining under incomplete evidence, AMIE）^[20]和规则挖掘系统 RuleN^[21]，后者包括神经逻辑编程（neural logic programming, Neural LP）^[22]和可微分规则挖掘（differentiable rule mining, DRUM）^[23]。相比于之前的方法，本文提出一种基于元学习的方法，挖掘和迁移与实体无关的元知识，从而解决分布式知识迁移问题。

1.3 联邦式图谱推理

联邦学习允许数据所有者将他们的私人数据存储在本地的客户端，并使用服务器协调客户端来训练一个全局模型。一般来说，联邦学习模型的训练是通过聚合本地计算的更新进行的，为了保护数据隐私，客户端的数据不会被转移到任何地方。最开始，联邦学习的提出是为了利用边缘设备中的数据，在云服务器上获得一个全局模型，即通过在边缘设备上利用数据进行本地训练，然后在云服务器上进行全局更新。McMahan H B 等人^[24]提出联邦平均（federated averaging, FedAvg）方法，试图通过几个客户端中的数据来学习一个共享模型。当前许多不同领域的研究人员都提出了多种解决方法，来完成在联邦设置下的不同任务。本文试图通过联邦学习的学习范式来学习多个具有重叠实体集合的知识图谱中的实体表示。

2 分布式知识迁移

定义单个知识图谱为 $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{P})$ ，其中 $\mathcal{E}$ 为实体集合， $\mathcal{R}$ 为关系集合， $\mathcal{P} = \{(h, r, t)\} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ 为三元组集合。在分布式知识迁移问题中，给定一

个源知识图谱的集合 $\mathcal{G}_s = \{\mathcal{G}_s^{(i)} = (\mathcal{E}_s^{(i)}, \mathcal{R}_s^{(i)}, \mathcal{P}_s^{(i)})\}_{i=1}^{n_s}$ ，以及目标知识图谱的集合 $\mathcal{G}_t = \{\mathcal{G}_t^{(i)} = (\mathcal{E}_t^{(i)}, \mathcal{R}_t^{(i)}, \mathcal{P}_t^{(i)})\}_{i=1}^{n_t}$ ，其中 n_s 和 n_t 为源知识图谱和目标知识图谱的数量，且源知识图谱和目标知识图谱之间的关系集合满足 $\bigcup_{i=1}^{n_t} \mathcal{R}_t^{(i)} \subseteq \bigcup_{i=1}^{n_s} \mathcal{R}_s^{(i)}$ 。为了行文的简洁，这里考虑 $n_s = n_t = 1$ 的情况，这并不会影响模型的设计及结果的验证。分布式知识迁移的目标是在 $\mathcal{G}_s$ 上学习一个模型，使得其可以泛化到 $\mathcal{G}_t$ 上，补全 $\mathcal{G}_t$ 中的知识图谱。传统的知识图谱表示学习模型对每一个实体学习一个表示，从而无法泛化到一个具有新实体的知识图谱上，例如从这里的源知识图谱到目标知识图谱。为了解决这个问题，需要设计一种无须对实体学习具体的表示的模型，该模型通过实体周围的结构模式对实体进行表示，这是因为相似的周围结构模式代表相似的语义信息。这种结构模式是通用的、可迁移的、与实体无关的，本文称之为“元知识”。为了解决分布式知识迁移问题，本文提出了基于元知识迁移的知识图谱嵌入模型（meta-knowledge transfer for knowledge graph embedding, MorsE）方法，使用该方法解决如下问题：①表示元知识；②在源知识图谱上学习元知识；③在目标知识图谱上适应元知识，从而完成知识图谱间的知识迁移。

2.1 元知识的表示

因为源知识图谱和目标知识图谱间具有重叠的关系，一个最自然的想法是迁移与关系有关的信息。因此，设计一个实体初始化器，利用每个实体周围所连接的关系，对实体的表示进行初始化。然而初始化得到的表示仅包含实体类型层面的信息，而不包含实例层面的信息。为了获得更好的实体表示，提出一个图神经网络调整器，通过节点的周围结构对初始化后的节点表示进行调整。在本节中，用一个具体的知识图谱 $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{P})$ 来描述如何对元知识进行表示。

在实体初始化器中，设计一个关系-域（relation-domain）表示 $\mathbf{R}^{\text{dom}} \in \mathbb{R}^{n_r \times d}$ 和关系-范围（relation-range）表示 $\mathbf{R}^{\text{ran}} \in \mathbb{R}^{n_r \times d}$ ，用于隐式地表示与某一种关系相连的头实体和尾实体的类型特征，其中 n_r 是关系的数量， d 是向量表示的维度。如图 1（a）所示，用下标 r 表示获取该矩阵的第 r 行，即某一个关系 r 的向量表示。对于一个

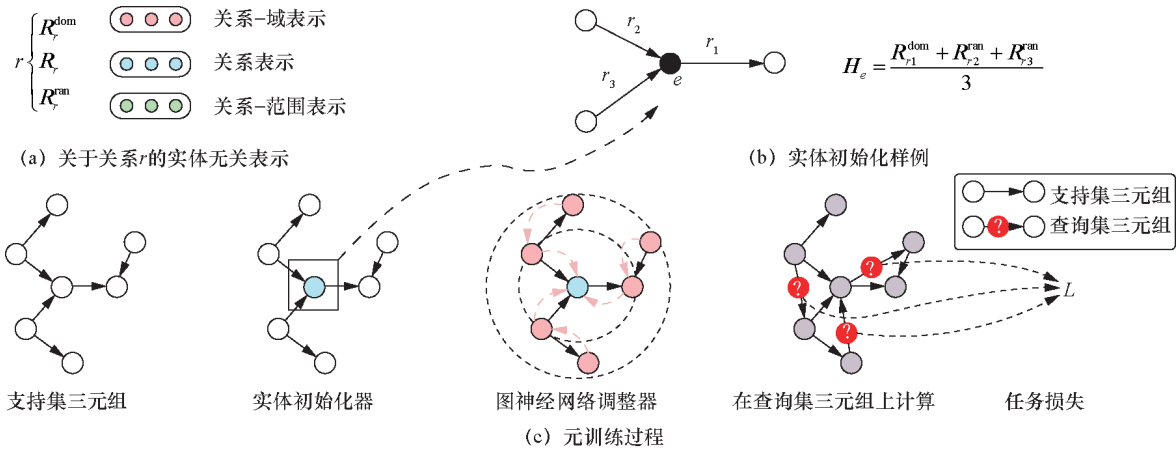


图1 分布式知识迁移概览

知识图谱, MorsE 首先利用实体连接的关系初始化实体, 具体来说, 对于一个实体 $e \in \mathcal{E}$, 其初始化的 H_e 如下:

$$H_e = \frac{\sum_{r \in \mathcal{O}(e)} R_r^{\text{dom}} + \sum_{r \in \mathcal{J}(e)} R_r^{\text{ran}}}{|\mathcal{O}(e)| + |\mathcal{J}(e)|} \quad (1)$$

其中, $\mathcal{J}(e) = \{r | \exists x, (x, r, e) \in \mathcal{P}\}$ 是进入实体 e 的关系集合, $\mathcal{O}(e)$ 是离开实体 e 的关系集合。实体初始化样例如图 1 (b) 所示。

在得到初始化的实体表示后, 利用图神经网络, 根据实体的周围结构对表示进行调整。这里的操作类似于关系图卷积网络 (relational graph convolutional network, R-GCN), 第 l 层的前向传播如下:

$$h_e^l = \alpha \left(\frac{1}{|\mathcal{N}(e)|} \sum_{(h,r) \in \mathcal{N}(e)} W_r^l h_h^{l-1} + W_0^l h_e^{l-1} \right) \quad (2)$$

其中, $\mathcal{N}(e) = \{(h, r) | (h, r, e) \in \mathcal{P}\}$ 是连接方向指向实体 e 的关系及相应的实体。 W_r^l 是第 l 层中关系 r 的关系特定转换矩阵。 W_0^l 是第 l 层中每一个关系的自连接转换矩阵。 h_e^l 是第 l 层中实体 e 的表示, 其中 $h_e^0 = H_e$ 。 $E_e = h_e^L$ 是最后一层的输出, 该输出是实体 e 的表示, L 表示图神经网络的总层数。

综上, 对于 $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{P})$, 元知识的表示过程为首先进行实体初始化:

$$H = \text{INIT}_{\theta}(\mathcal{P}) \quad (3)$$

其中, θ 是包含 R^{dom} 和 R^{ran} 的参数。然后利用上述多层图神经网络的前向传播, 将表示调整为:

$$E = \text{MODULATE}_{\phi}(\mathcal{P}, H) \quad (4)$$

其中, ϕ 包含图神经网络调整器中的所有参数。整

个过程可以看作:

$$E = \text{REP}_{\theta, \phi}(\mathcal{P}) = \text{MODULATE}_{\phi}(\mathcal{P}, \text{INIT}_{\theta}(\mathcal{P})) \quad (5)$$

2.2 元知识的学习

第 2.1 节对元知识进行表示和建模, 并且利用实体的周围结构模式对实体进行表示, 但为了让最终输出的实体表示是有意义的, 需要对元学习表示过程中的参数进行学习。为了解决元知识的学习, 这里利用元学习的方法, 也就是“学习如何学习”。元学习的方法在一系列任务而非数据点上训练, 其中每个任务包含自己的训练集和测试集, 这里称之为支持集和查询集。根据以往的元学习任务设定, 在 MorsE 中, 设定元学习任务, 从而模拟得到实体表示, 再验证实体表示的过程。具体来说, 为了在源知识图谱上学习元知识, 从 $\mathcal{G}_s$ 上随机采样子图谱 $\{\mathcal{G}_{Si} = (\mathcal{E}_{Si}, \mathcal{R}_{Si}, \mathcal{P}_{Si})\}_{i=1}^m$ 从而形成训练任务, 其中一个任务包含自己的支持集和查询集, 如 $T_i = (\mathcal{S}_i, \mathcal{Q}_i)$ 。具体来说, 从子图谱中构建任务的方式如下:

$$T_i = (\mathcal{S}_i = \{(h, r, t) \sim \mathcal{P}_{Si}\}, \mathcal{Q}_i = \{(s, p, o) \sim \mathcal{P}_{Si}\}) \quad (6)$$

$$\text{s.t. } \mathcal{S}_i \cap \mathcal{Q}_i = \emptyset, \{p\} \subseteq \{r\}$$

$$\{\{s\} \cup \{o\}\} \subseteq \{\{h\} \cup \{t\}\}$$

其中支持集和查询集中的三元组互斥, 并且在查询集中出现的实体、关系同时需要在支持集中也出现。这样的任务设定是为了让模型可以获得任意知识图谱的实体表示, 而非具体某一个固定实体集合的知识图谱的实体表示。

在上述元学习任务的基础上, 对元知识进行学

习，该过程也被叫作元训练过程。在元训练过程中，在支持集上获得实体的表示，在查询集上进行验证。整个元训练的训练目标如下：

$$\min_{\theta} \sum_{i=1}^m \mathcal{L}(Q_i, \text{REP}_{\theta, \phi}(\mathcal{S}_i), \mathbf{R}) \quad (7)$$

其中， $\mathbf{R} \in \mathbb{R}^{n_r \times d}$ 是关系表示矩阵， $\{\phi, \theta, \mathbf{R}\}$ 是可学习的参数。元训练过程如图 1 (c) 所示。

在具体介绍损失函数 $\mathcal{L}$ 的计算方式之前，先介绍如何利用实体和关系表示计算一个三元组 (h, r, t) 的真值，即计算打分函数 $s(h, r, t)$ 。不同的知识图谱表示学习方法会设计不同的打分函数，这里使用 4 种具有代表性的知识图谱表示方法，分别是建模多关系数据的翻译嵌入 (translating embeddings for modeling multi-relational data, TransE) 模型、距离乘法 (distance multiplication, DistMult) 模型、复数嵌入 (complex embeddings, ComplEx) 模型以及基于复数空间中的关系旋转 (relational rotation in complex space, RotatE) 模型。不同知识图谱表示学习方法的打分函数见表 1，其中 $\text{Re}(\cdot)$ 表示复数的实数部分， $\circ$ 表示向量的对应元素相乘， $\mathbf{h}, \mathbf{r}$ 和 $\mathbf{t}$ 分别表示三元组 (h, r, t) 各部分的向量表示，实体的向量由 $\text{REP}_{\theta, \phi}(\cdot)$ 计算，关系的向量保存在可训练参数 $\mathbf{R}$ 中。

表 1 不同知识图谱表示学习方法的打分函数

知识图谱表示学习方法	打分函数	向量空间
TransE	$-\ \mathbf{h} + \mathbf{r} - \mathbf{t}\ $	$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^d$
DistMult	$\mathbf{h}^\top \text{diag}(\mathbf{r}) \mathbf{t}$	$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^d$
ComplEx	$\text{Re}(\mathbf{h}^\top \text{diag}(\mathbf{r}) \bar{\mathbf{t}})$	$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{C}^d$
RotatE	$-\ \mathbf{h} \circ \mathbf{r} - \mathbf{t}\ $	$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{C}^d$

在第 i 个任务上的损失函数如下：

$$\mathcal{L} = -\sum_{(h, r, t) \in Q_i} \log \sigma(s(h, r, t)) - \sum_{i=1}^k p(h'_i, r, t'_i) \log \sigma(-s(h'_i, r, t'_i)) \quad (8)$$

其中， σ 是 Sigmoid 函数， k 是负样本的个数， (h'_i, r, t'_i) 是第 i 个负样本， $p(h'_i, r, t'_i)$ 是自对抗负采样的权重，计算式如下：

$$p(h'_j, r, t'_j) = \frac{\exp \beta s(h'_j, r, t'_j)}{\sum_i \exp \beta s(h'_i, r, t'_i)} \quad (9)$$

其中， β 是 Softmax 函数中的温度参数。

2.3 元知识的适应

在源知识图谱上进行元训练，学习得到参数以

及元知识后，考虑如何将学习好的参数应用到目标知识图谱上，从而完成元知识的迁移。得益于已经学习好的参数，对于任何一个目标知识图谱 $\mathcal{G}_t = (\mathcal{E}_t, \mathcal{R}_t, \mathcal{P}_t)$ ，都可以通过 $\text{REP}_{\theta, \phi}(\mathcal{P}_t)$ 获得该目标知识图谱的实体表示，从而将元知识在目标知识图谱上进行适应，完成从源知识图谱到目标知识图谱的知识迁移。

3 联邦式图谱推理

相比于上述分布式知识迁移对于单个知识图谱的定义，联邦式知识图谱定义一个知识图谱集合为 $\mathcal{G}^F = \{\mathcal{G}_i\}_{i=1}^C$ ，其中 $\mathcal{G}_i = \{\mathcal{E}_i, \mathcal{R}_i, \mathcal{P}_i\}$ 是一个存储在客户端 i 上的知识图谱， C 是客户端的总数，不同客户端的知识图谱之间具有重叠的实体集合。除此之外，还有一个服务器端，用来保存实体表 T ，该表存储了不同客户端中的所有实体，并对不同客户端之间的实体进行对应。

在联邦式图谱推理中，笔者关心的是如何对联邦知识图谱进行补全。该补全任务主要关注如何预测每个客户端缺失的三元组，如 $(h, r, ?)$ ，并且在保证数据隐私的前提下同时利用其他知识图谱中的信息。

为了解决联邦知识图谱补全问题，这里提出一种联邦知识图谱嵌入 (federated knowledge graph embedding, FedE) 框架。仿照先前提出的联邦学习算法 FedAvg，笔者将 FedE 建模成两个部分：一个服务器端，多个客户端。不同的知识图谱被放置在多个客户端上，不同的知识图谱之间有重叠的实体集合，并且这些不同的知识图谱会定义自己的三元组和关系集合。不同客户端上的知识图谱使用它们自己的三元组，通过客户端更新对实体、关系的嵌入表示进行更新。服务器端通过服务器操作负责控制实体嵌入表示的聚合和共享。服务器端上会存储并维护一个实体表格，用于记录所有客户端中的实体，并且负责将不同客户端知识图谱中的实体对应到服务器端的实体表格上。

首先介绍服务器操作和客户端更新两个流程，然后介绍模型融合的过程。在模型融合中，笔者对从联邦知识图谱学习而来的嵌入表示和从客户端自身三元组学习而来的嵌入表示进行模型融合，从而从不同角度提升嵌入表示的能力。联邦式图谱推理概览如图 2 所示。

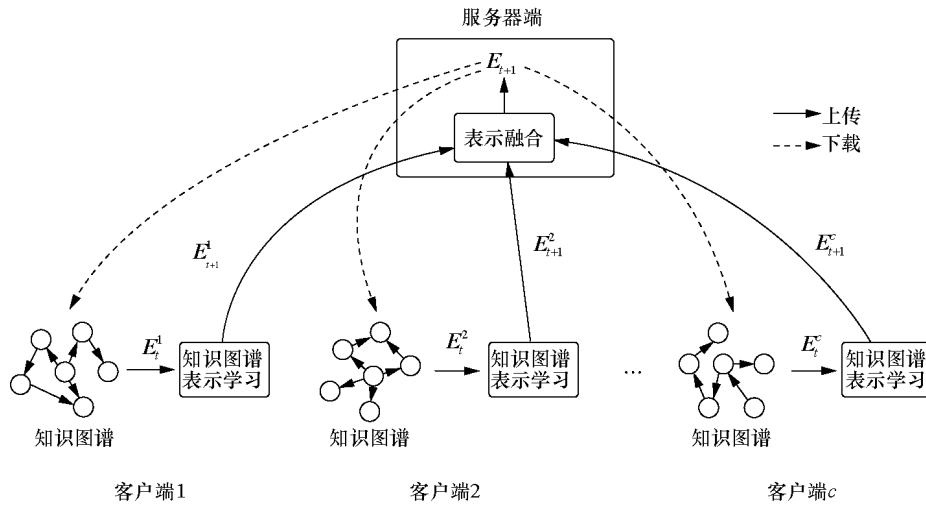


图2 联邦式图谱推理概览

3.1 服务器操作

在笔者提出的 FedE 框架中，服务器端负责聚合来自不同客户端的实体嵌入表示，并将聚合的实体嵌入表示发送回每个客户端。笔者将单个服务器端聚合操作称为 FedE 的一轮操作。

首先，基于实体表 T ，服务器端构造一组置换矩阵 $\{P^c \in \{0,1\}^{n \times n_c}\}_{c=1}^C$ 和存在向量 $\{v^c \in \{0,1\}^{n \times 1}\}_{c=1}^C$ ，用于聚合来自客户端的实体嵌入表示，其中 C 是客户端的总数， n 是实体表中所有实体的数量， n_c 是来自客户端 c 的实体数量。具体来说，置换矩阵 P^c 用于将客户端 c 的实体嵌入表示矩阵映射到服务端的实体表，其中如果实体表中的第 i 个实体对应于客户端 c 中第 j 个实体，那么 $P_{i,j}^c = 1$ 。 $v_i^c = 1$ 表示实体表中的第 i 个实体（即 T_i ）存在于客户端 c 中。

其次，服务器会根据实体表 T 随机初始化实体嵌入表示矩阵 $E_0 \in \mathbb{R}^{n \times d_e}$ ，其中 d_e 是实体嵌入表示的维度， n 是 T 中所有实体的数量。在向每个客户端发送实体嵌入表示之前，进行一次矩阵排列转换。具体来说，对于客户端 c ，进行如下操作：

$$E_0^c \leftarrow P^{c\top} E_0 \quad (10)$$

然后，模型进行多轮训练。在第 t 轮中， $F \times C$ 个客户端将被随机选择并作为集合 C_t ，其中 $F \in [0,1]$ 是被选择的客户端的比例。在每个客户端 $c \in C_t$ 中，对实体嵌入表示矩阵训练特定周期后，服务器会聚合来自客户端的实体嵌入表示矩阵 E_{t+1}^c ：

$$E_{t+1} \leftarrow \left(\mathbf{1} \oslash \sum_{c \in C_t} v^c \right) \otimes \sum_{c \in C_t} P^c E_{t+1}^c \quad (11)$$

其中， $\mathbf{1}$ 表示一个全为 1 的向量， $\oslash$ 表示向量间的按位除法， $\otimes$ 表示具有广播操作的向量间按位乘法，也就是 $[v \otimes M]_{i,j} = v_i \times M_{i,j}$ 。

在 t 轮训练结束时，在服务器端聚合完成的 E_{t+1} 将按照矩阵排列转换操作进行重新排列，然后分发回每个客户端 $c \in C_t$ ，即每个客户端的实体嵌入表示矩阵 E^c 将由 $P^{c\top} E_{t+1}$ 进行更新。

3.2 客户端更新

在第 t 轮训练中，客户端 c 根据特定的知识图谱嵌入方法更新实体嵌入表示矩阵 E^c 和关系嵌入表示矩阵 R^c 。

对于客户端 c 中的三元组 $(h, r, t) \in \mathcal{G}_c$ ，通过表 1 中的相应知识图谱表示学习方法的打分函数 $s(h, r, t)$ 计算三元组的似然性。依据知识图谱的自对抗损失函数对客户端中的实体和关系表示进行更新。客户端 c 在第 t 轮训练了 P 个周期后，生成新的实体嵌入表示矩阵 E_{t+1}^c ，并将其上传到服务器中。

3.3 模型融合

服务器操作和客户端更新的过程一直迭代进行，直到收敛。然而，学习到的知识图谱嵌入表示是基于所有客户端的全局最小值进行优化的，而不是每个客户端的局部最小值。为此，笔者设计了一个模型融合过程来集成从本地知识图谱和联邦知识图谱中学习到的知识图谱嵌入表示。

每个客户端完成了基于联邦知识图谱的嵌入表示学习和基于本地知识图谱的嵌入表示学习后，再进行模型融合操作。具体而言，基于联邦知识图谱的嵌入表示按照服务器操作和客户端更新迭代

进行训练，基于本地知识图谱的嵌入表示使用客户端更新过程中的自对抗损失函数为每个客户端单独训练嵌入表示。

对于三元组 (h, r, t) ，首先将分数 $s^l(h, r, t)$ 和 $s^f(h, r, t)$ 连接起来作为特征向量，其中由 l 作为上标的打分函数使用由本地知识图谱训练而来的嵌入表示，由 f 作为上标的打分函数使用由联邦知识图谱训练而来的嵌入表示。然后，将特征向量作为输入，使用线性分类器输出这个三元组的最终分数 $s(h, r, t)$ ：

$$s(h, r, t) = Wx + b \quad (12)$$

$$x = [s^l(h, r, t); s^f(h, r, t)] \quad (13)$$

其中， W 和 b 是可训练的模型参数。

这个线性分类器使用边距排序损失训练使正样本三元组的排名高于负样本三元组的排名， (h, r, t) 的模型融合损失定义如下：

$$\mathcal{L}_f(h, r, t) = \max(0, \gamma - s(h, r, t) + s(h, r, t')) \quad (14)$$

其中， γ 是边距， (h, r, t') 是负三元组。模型融合过程的训练目标是 최소화 所有三元组的模型融合损失。

3.4 隐私保护

本节从 3 个层次对 FedE 的隐私保护进行讨论。

① 关系层次。关系及其关系嵌入表示不会离开它们所存在的客户端，每个客户端都不知道其他客户端的关系，因此可以保护关系的隐私。

② 实体层次。通过加密领域的现有研究，私有集合交集方法和秘密共享方法可以安全地获取和存储服务器中的实体表。客户端之间的唯一连接是对齐实体的嵌入表示。然而，对于不同客户端的相同实体，即使它们最终具有相同的嵌入表示，客户端的所有者也无法解码其他客户端是否具有特定实体，因为实体嵌入仅隐式地包含相关语义信息（即实体的邻居结构），所以可以保护实体的隐私。

③ 三元组层次。在 FedE 中，不同客户端的三元组不会被转移，这直接保证了三元组的数据隐私。此外，在预测一个客户端的两个实体之间的关系时，由于客户端不知道其他客户端的关系，客户端所有者无法推断出其他客户端中的三元组。然而，如果两个客户端有两个语义相似的关系，并且有大量相同的三元组，那么这两个关系有可能具有相似的嵌入，一个客户端将预测出与另一个客户端

相似的三元组。由于客户不知道彼此的关系，因此此类数据泄露的可能性极小。最后，如果需要严格地保护三元组的隐私，笔者建议在进行 FedE 时使用来自不同域（具有完全不同的关系）的客户端来增强彼此的知识图谱表示能力。

4 实验分析

本节分别对分布式知识迁移和联邦式图谱推理的效果进行验证。

4.1 分布式知识迁移

对于知识图谱链接预测的分布式知识迁移任务，MorsE 在源知识图谱的三元组上训练，然后在目标知识图谱的三元组上进行模型适应，最后根据目标知识图谱的链接预测效果进行模型评估。

本文使用的数据集来源于原始的 WN18RR^[25]、FB15k-237^[26]和 NELL-995^[27]数据集，经过 Teru K 等人^[17]的处理后，这些数据集可以被用于对归纳关系预测任务的评测，该任务对应于本文所提出的分布式知识迁移。每个数据集都被采样为 4 个不同的版本（v1~v4），不同的版本具有不同的规模，以进行鲁棒的评估。具体来说，每个数据集有两部分，对应分布式知识迁移任务中的源知识图谱和目标知识图谱。

对于评估指标，笔者最终使用 Hits@10 指标评估目标知识图谱中链接预测的性能，Hist@10 表示正确的三元组在所有候选三元组中排序进入前 10 的比例。所有结果皆为头实体和尾实体链接预测结果的平均值。遵循基线模型，通过将每个测试三元组在 50 个其他随机采样的候选负三元组中排名 5 次来近似链接预测评估结果。

将 MorsE 与两类可以进行分布式知识迁移完成链接预测的基线方法（分别是基于规则学习的方法和基于 GraIL 的方法）进行对比。基于规则学习的方法包括显示地从知识图谱中提取规则的 RuleN，以及 Neural LP 和 DRUM，它们以端到端的方式学习规则。基于 GraIL 的方法包括 GraIL 及其改进方法 CoMPILE，它们在子图上学习图神经网络，并进行关系预测，从而处理新出现的实体。

表 2 展示了本文提出的方法在使用不同知识图谱表示学习模型的打分函数时的结果。其中，带下划线的数字代表基线模型中最好的结果，加粗的数字代表所有模型中最好的结果。从表 2 中

表2 分布式知识迁移实验结果

模型名称	WN18RR				FB15k-237				NELL-995			
	v1	v2	v3	v4	v1	v2	v3	v4	v1	v2	v3	v4
Neural LP	74.4	69.0	46.2	67.1	52.9	58.9	52.9	55.9	40.8	78.7	82.7	<u>80.6</u>
DRUM	74.4	69.0	46.2	67.1	52.9	58.7	52.9	55.9	19.4	78.6	82.7	<u>80.6</u>
RuleN	80.9	78.2	53.4	71.6	49.8	77.8	<u>87.7</u>	85.6	53.5	81.8	77.3	61.4
GraIL	82.5	78.7	58.4	73.4	64.1	81.8	82.8	<u>89.3</u>	<u>59.5</u>	93.3	91.4	73.1
CoMPILE	<u>83.6</u>	<u>79.8</u>	<u>60.7</u>	<u>75.5</u>	<u>67.6</u>	<u>83.0</u>	84.7	87.4	58.4	93.9	92.8	75.1
MorsE (TransE)	81.5	78.8	67.3	76.5	84.7	96.3	95.8	96.4	51.8	83.8	89.3	82.4
MorsE (DistMult)	83.7	80.0	66.1	79.3	82.5	94.9	94.9	95.8	66.6	79.4	84.3	54.1
MorsE (ComplEx)	83.5	81.4	67.3	79.4	80.0	92.8	93.6	95.1	65.0	76.9	82.3	42.6
MorsE (RotatE)	84.1	81.5	70.9	79.6	83.2	95.7	95.7	95.9	65.2	80.7	87.7	53.4

可以发现, 本文提出的模型在大部分的数据集上都能得到最好的效果。尽管本文只对最好的结果进行了加粗, 但使用不同打分函数的 MorsE 方法相比于大多基线方法可以获得稳定的性能提升。例如, 在 FB15k-237 的 4 个数据集上, 分别使用 TransE、DistMult、ComplEx 以及 RotatE 的 MorsE 模型相比于基线方法中最好结果, 平均获得了 14.6%、13.0%、10.8% 以及 13.7% 的提升。以上观察结果说明了 MorsE 对分布式知识迁移完成知识图谱链接预测的有效性。

4.2 联邦式图谱推理

在联邦式图谱推理任务上, 为了测试 FedE 模型完成联邦式图谱推理的有效性, 本文在知识图谱基准数据集 FB15k-237 进行验证。该基准数据集最初是为了评估知识图嵌入方法而提出的。为了在联邦设置中评估 FedE, 笔者将三元组随机分布到不同的客户端中。具体方法是, 根据联邦式知识图谱的定义, 随机选择关系并放入不同客户端, 并根据选择的关系将三元组分配到客户端。将 FB15k-237 随机分为 3 个客户端, 获得了 FB15k-237-Fed3 数据集。训练、验证、测试三元组的比例为 80%、10%、10%。

遵循知识图嵌入方法的传统评估指标, 利用平均倒数排名 (mean reciprocal rank, MRR) 和 Hits@10, 评估在每个客户端中进行链接预测的性

能。链接预测的结果是在过滤设置下进行计算的, 即在进行结果排序时, 过滤并删除出现在训练、验证和测试集中的候选三元组。

为了展示 FedE 框架的能力, 在独立 (isolation) 和聚集 (collection) 两种设置下对比 FedE 和传统的知识图谱表示学习方法。对于 isolation 设置, 知识图嵌入仅针对每个客户端在其自己的知识图谱上进行训练。对于 collection 设置, 将来自不同客户端的知识图谱三元组收集在一起, 并在整个三元组上训练知识图嵌入 (请注意, 此 collection 设置违反了隐私保护, 这是因为在收集过程中会暴露来自不同客户端的三元组)。对于 FedE 设置, 在联邦设定中使用本文提出的 FedE 训练知识图嵌入, 并且每个客户端上的三元组不会被转移。本文选择 4 种典型的知识图表示学习方法 TransE、DistMult、ComplEx 和 RotatE, 用于设置 isolation、collection、FedE。

联邦式图谱推理实验结果见表 3, 带下划线的数据表示 FedE 和 isolation 两种设置下最好的结果, 加粗数据表示在所有设置下最好的结果。基于这些结果可以发现, 本文提出的 FedE 比在 isolation 设置下训练的知识图嵌入获得的链接预测性能更好。举例来说, 将 RotatE 作为知识图谱表示学习方法, FedE 在 FB15k-237-Fed3 上的结果显示, MRR 和 Hits@10 分别相对提高了 5.8% 和 7.2%。这些结果显示了 FedE 的有效性。

表 3 联邦式图谱推理实验结果

知识图谱表示学习方法	设置	客户端 1		客户端 2		客户端 3	
		MRR	Hits@10	MRR	Hits@10	MRR	Hits@10
TransE	isolation	50.4	69.8	38.7	63.7	33.7	59.7
	collection	54.2	74.1	41.0	66.9	35.7	63.1
	FedE	54.8	74.8	41.8	68.4	36.3	63.4
DistMult	isolation	48.9	65.2	36.6	56.2	34.7	57.0
	collection	51.7	70.1	38.2	59.9	33.9	56.4
	FedE	53.6	70.3	40.4	62.4	37.5	61.5
ComplEx	isolation	48.4	65.3	37.2	57.8	32.6	55.7
	collection	50.0	67.5	36.9	57.4	33.4	55.5
	FedE	52.9	70.2	40.9	63.2	36.1	59.5
RotatE	isolation	52.7	70.6	40.6	64.2	36.6	60.2
	collection	55.7	75.1	42.1	67.3	37.1	63.5
	FedE	55.5	75.4	43.0	68.5	38.9	65.0

5 结束语

本文对群体知识图谱面临的分布式知识迁移和联邦式图谱推理问题进行模型设计和实验验证。在分布式知识迁移中，利用元知识迁移建模并迁移与实体无关的元知识，从而有效解决单个知识图谱不完备的问题。在联邦式图谱推理中，基于联邦学习框架在不同客户端上学习知识图谱表示，并在服务器端上进行聚合。未来笔者会探讨更加复杂的群体知识图谱场景，例如在分布式知识迁移的场景下，考虑新增的关系以及在联邦式图谱推理中考虑新增客户端会带来新增实体和关系的情况。

参考文献：

- [1] BERNERS-LEE T, HENDLER J, LASSILA O. The semantic Web[J]. Scientific American, 2001, 284(5): 34-43.
- [2] VRANDEČIĆ D, KRÖTZSCH M. Wikidata[J]. Communications of the ACM, 2014, 57(10): 78-85.
- [3] AUER S, BIZER C, KOBILAROV G, et al. DBpedia: a nucleus for a web of open data[M]//The semantic Web. Heidelberg: Springer, 2007.
- [4] MITCHELL T, COHEN W, HRUSCHKA E, et al. Never-ending learning[J]. Communications of the ACM, 2018, 61(5): 103-115.
- [5] XIONG W H, DU J F, WANG W Y, et al. Pretrained encyclopedia: weakly supervised knowledge-pretrained language model[C]//Proceedings of the International Conference on Learning Representations. [S.l.:s.n.], 2019.
- [6] YASUNAGA M, REN H Y, BOSSELU T A, et al. QA-GNN: reasoning with language models and knowledge graphs for question answer-

ing[C]//Proceedings of 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2021: 535-546.

- [7] ZHANG Q, ZHANG L, QIN C, et al. A survey on knowledge graph-based recommender systems[J]. Scientia Sinica Informationis, 2020, 50(7): 937-956.
- [8] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, 2018, arXiv:1810.04805.
- [9] WANG X Z, GAO T Y, ZHU Z C, et al. KEPLER: a unified model for knowledge embedding and pre-trained language representation[J]. Transactions of the Association for Computational Linguistics, 2021, 9: 176-194.
- [10] CHEN H J, HU N, QI G L, et al. OpenKG chain: a blockchain infrastructure for open knowledge graphs[J]. Data Intelligence, 2021: 1-18.
- [11] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[J]. Advances in Neural Information Processing Systems, 2013, 26.
- [12] WANG Z, ZHANG J W, FENG J L, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.:s.n.], 2014.
- [13] ZHOU X H, YI Y H, JIA G. Path-RotatE: knowledge graph embedding by relational rotation of path in complex space[C]//Proceedings of 2021 IEEE/CIC International Conference on Communications in China. Piscataway: IEEE Press, 2021: 905-910.
- [14] NICKEL M, TRESP V, KRIEGER H. A three-way model for collective learning on multi-relational data[C]//Proceedings of the 28th International Conference on International Conference on Machine Learning. [S.l.:s.n.], 2011: 809-816.
- [15] YANG B S, YIH W T, HE X D, et al. Embedding entities and relations

- for learning and inference in knowledge bases[J]. arXiv preprint, 2014, arXiv:1412.6575.
- [16] TROUILLON T, WELBL J, Riedel S R, et al. Complex embeddings for simple link prediction[C]//Proceedings of the International Conference on Machine Learning. [S.l.:s.n.], 2016: 2071-2080.
- [17] TERU K, DENIS E, HAMILTON W. Inductive relation prediction by subgraph reasoning[C]//Proceedings of the International Conference on Machine Learning. [S.l.:s.n.], 2020: 9448-9457.
- [18] MAI S J, ZHENG S J, YANG Y D, et al. Communicative message passing for inductive relation reasoning[J]. arXiv preprint, 2020, arXiv:2012.08911.
- [19] CHEN J J, HE H R, WU F, et al. Topology-aware correlations between relations for inductive link prediction in knowledge graphs[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.:s.n.], 2021: 6271-6278.
- [20] GALÁRRAGA L A, TEFLIOUDI C, HOSE K, et al. AMIE: association rule mining under incomplete evidence in ontological knowledge bases[C]//Proceedings of the 22nd International Conference on World Wide Web. [S.l.:s.n.], 2013: 413-422.
- [21] MEILICKE C, FINK M, WANG Y J, et al. Fine-grained evaluation of rule- and embedding-based systems for knowledge graph completion[C]//Proceedings of the International Semantic Web Conference. [S.l.:s.n.], 2018: 3-20.
- [22] YANG F, YANG Z L, COHEN W W. Differentiable learning of logical rules for knowledge base reasoning[J]. arXiv preprint, 2017, arXiv:1702.08367.
- [23] SADEGHIAN A, ARMANDPOUR M, DING P, et al. DRUM: end-to-end differentiable rule mining on knowledge graphs[J]. Advances in Neural Information Processing Systems, 2019, 32: 15347-15357.
- [24] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[J]. arXiv preprint, 2016, arXiv:1602.05629.
- [25] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2D knowledge graph embeddings[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.:s.n.], 2018.
- [26] TOUTANOVA K, CHEN D Q, PANTEL P, et al. Representing text for joint embedding of text and knowledge bases[C]//Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 1499-1509.
- [27] XIONG W H, HOANG T, WANG W Y. DeepPath: a reinforcement learning method for knowledge graph reasoning[C]//Proceedings of 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2017: 564-573.

[作者简介]



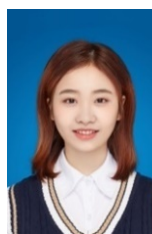
陈名杨（1997—），男，浙江大学计算机科学与技术学院博士生，主要研究方向为知识图谱表示学习。



张文（1992—），女，浙江大学软件学院助理研究员，主要研究方向为知识图谱、大数据系统、图数据处理。



陈湘楠（1998—），男，浙江大学软件学院硕士生，主要研究方向为知识图谱表示学习。



周虹廷（1997—），女，浙江大学计算机科学与技术学院硕士生，主要研究方向为知识图谱和图神经网络。



陈华钧（1978—），男，博士，浙江大学计算机科学与技术学院教授，主要研究方向为知识图谱、大数据系统与自然语言处理。