

基于分布式表示的知识图谱推理研究

陈光希¹ 冯 峰²

CHEN Guangxi FENG Feng

摘 要

知识图谱推理 (knowledge graph reasoning, KGR) 是知识图谱构建及相关应用领域的一个热点问题, 其目的是利用知识图谱中已经存在的信息来推导出新的知识。为了更好地挖掘和利用知识图谱中的隐含信息, 基于分布式表示的知识图谱推理方法应运而生, 文章从 KGR 的定义入手, 对现有的基于分布式表示的 KGR 方法进行分类分析。基于分布式表示的 KGR 方法可以主要分为基于张量分解和基于翻译模型的方法, 前者通过将知识图谱表示为张量并进行分解来捕捉复杂关系, 后者通过在低维空间中进行平移或旋转操作来表示实体和关系。与此同时, 文章探讨了所研究方法在实际应用中的效果, 指出张量分解方法在表达复杂关系上所具有的优势, 但计算复杂且扩展性差; 翻译模型计算高效, 但在处理复杂关系时存在局限。最后, 展望了该领域的未来研究方向, 指出了当前研究中存在的挑战和解决的途径, 旨在为研究人员和应用开发者提供有价值的参考。

关键词

知识图谱推理; 分布式表示; 张量分解; 翻译模型; 人工智能

doi: 10.3969/j.issn.1672-9528.2025.03.040

0 引言

知识图谱 (knowledge graph, KG) 的概念最早由 Google 在 2012 年提出, 被定义为由大量实体及其之间关系组成的大规模知识库。近年来, 知识图谱作为一种语义网络, 在自然语言处理^[1]、智能问答系统^[2]、智能推荐等领域都得到了广泛应用。尽管 Freebase^[3]、DBpedia^[4]、YAGO^[5]、NELL^[6] 等常用的大规模知识图谱包含了数以百万计的实体和关系, 但内部许多实体之间的隐含关系还未被充分挖掘。知识图谱逐渐成为研究的热点。

知识图谱推理, 或称知识推理, 指在既定的知识图谱中通过推理技术推导出实体间潜在或者新的关系, 发现新的知识。通过不断优化和应用知识图谱推理技术, 可以更高效地挖掘知识图谱中的信息, 为各种应用场景提供更为智能和高效的解决方案。因此, 本文提出了知识图谱推理研究, 通过预测现有实体之间的潜在关系, 推理出新的关系事实, 扩大知识图谱的规模。

随着深度学习和表示学习技术的发展, 基于分布式表示的知识图谱推理方法成为该领域的研究热点。分布式表示将知识图谱中的实体和关系嵌入到低维向量空间中, 通过在这个向量空间中进行计算和推理, 可以更高效地发现实体间的潜在关系。相较于传统的符号表示方法, 分布式表示具有处

理大规模数据、捕捉复杂关系和泛化能力强等优点。基于分布式表示的知识图谱推理方法主要分为两类: 基于张量分解的方法和基于翻译模型的方法。本文将详细综述这两种方法, 分析其优缺点和应用场景, 并探讨基于分布式表示的知识图谱推理在实际应用中的效果和表现。此外, 本文还展望了其未来研究方向, 为后续研究提供有价值的参考。

1 知识图谱推理概述

由于大多数知识图谱都是手工或者半自动构造的, 大量隐含的实体和关系都未被发现, 因此不完备性是几乎所有知识图谱的普遍问题。知识图谱推理 (knowledge graph reasoning, KGR) 的目标是通过分析和计算知识图谱中实体及其关系的结构和属性, 来识别和预测实体间的潜在关系, 从而发现隐含的知识, 提升知识图谱的完备性和准确性, 进而增强其在各种应用场景中的实用性。在知识推理中, 知识图谱提供了对人类知识和经验的总结, 而推理技术则基于知识图谱中已有的知识, 实现了对潜在知识和未知知识的发现, 拓展了知识问答、个性化搜索^[7]、智能推荐^[8]等能力。

早期的推理研究是由逻辑学和知识工程领域的学者进行的。逻辑学者主张用形式化的方法来描述客观世界, 认为一切推理都是基于现有的逻辑知识, 如一阶逻辑和谓词推理^[9]。关注点为如何从已知的命题和谓词中得到正确的结论。与逻辑领域的学者使用命题或者一阶谓词来表示客观世界中的概念不同, 知识工程领域的学者使用语义网络来表示更丰富的概念和知识来描述实体和属性之间的关系。

1. 北京无线电计量测试研究所 北京 100854

2. 山东省网络安全与信息化技术中心 山东济南 250002

随着互联网数据规模的爆炸式增长，基于人工构建知识库的传统方法已无法适应大数据时代对海量知识的需求，因此，数据驱动的机器推理方法逐渐成为知识推理研究的主流。数据驱动的方法通过从大规模数据集中自动提取和学习知识，不仅提升了知识库的构建效率，还能动态更新和扩展知识图谱。然而，这些方法也面临如何有效表示和推理复杂关系的问题。

为了更高效地处理和推理知识图谱中的复杂关系，基于分布式表示的知识图谱推理方法应运而生。分布式表示（distributed representation）通过将知识图谱中的实体和关系嵌入到低维向量空间中，使得复杂的图结构可以通过向量运算进行处理，从而大大提升了推理的效率和精度。基于分布式表示学习的方法能够捕捉实体和关系间的潜在模式，并通过向量表示简化推理过程，逐渐成为知识图谱推理研究的主流方向。

具体而言，基于分布式表示的知识图谱推理方法可以分为基于张量分解和基于翻译模型两大类方法。

2 方法

本节将从张量分解和翻译模型两类来介绍基于分布式表示的知识图谱推理方法的发展。图 1 总结了基于分布式表示的知识图谱推理方法的时间轴。

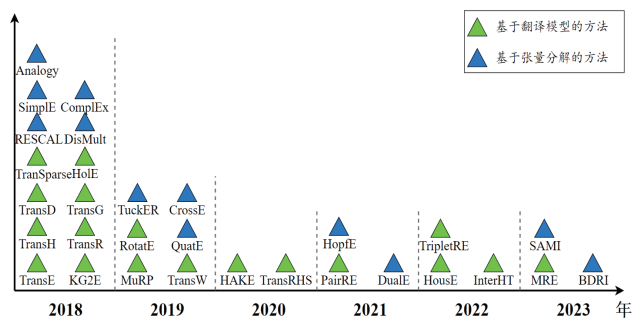


图 1 基于分布式表示的知识图谱推理方法时间轴

2.1 基于张量分解的模型

张量分解模型将 KG 编码为向量，进而分解为实体和关系的低维向量组合。

RESICAL^[10] 作为第一个张量分解模型，是一种基于三元张量分解的关系学习方法，其中每个实体都与一个向量相关联。如图 2 所示，RESICAL 使用向量来捕获每个实体的潜在语义，并进一步利用矩阵将潜在因素之间的相互作用建模为矩阵。然而，RESICAL 模型参数过于复杂，为了简化，Yang 等人提出了如图 3 所示的 DisMult 模型，该模型将关系矩阵分解为对角矩阵，来减少每个关系的参数。以前的知识图谱推理模型使用三维二元张量来进行链接预测，张量的每个切片表示知识图谱中某一关系的邻接矩阵。通过对张量进行低

秩分解和多线性计算来进行链接预测。然而这种方法不太适用于非堆成关系的预测。因此 ComplEx 模型通过使用复值嵌入来扩展 DisMult 模型，从而改进了不对称关系的建模。此外，HolE 通过结合 RECAL 的表达能力和 DisMult 的效率和简单性来学习整个知识图谱的向量空间表示。HolE 可以捕获丰富的交互，且仍然可以高效地计算并轻松扩展到非常大的数据集上。Analogy 设计了具有类比结构约束的双线性评分函数用于捕获实体之间丰富的交互，填补了从类推推理的角度预测新三元组的空白。

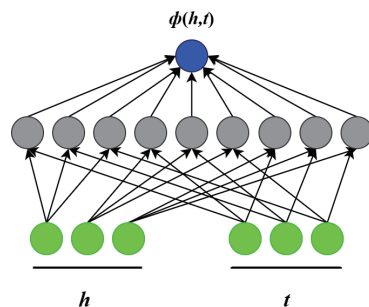


图 2 RESCAL 模型

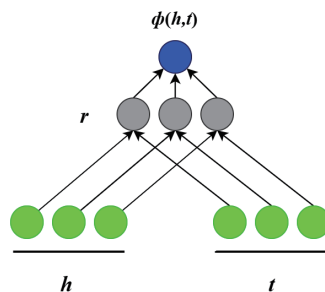


图 3 DisMult 模型

此外，一些模型也开始尝试替代分解操作。Simple 增加了两个独立实体嵌入的正则 CP 分解。TuckER 首次适用 Tucker 分解。CrossE 通过特定于关系的交互矩阵考虑实体之间的交叉交互。QuatE 基于关系旋转四元组表示实现了头尾之间的语义匹配。受其启发，DualE 将嵌入投影到对偶四元组空间中，以实现平移和旋转操作的统一框架。此外，HopfE 在不失可解释性的前提下，利用了四维超球空间的结构属性和语义属性。此外，QuatRE 通过汉密尔顿积增强实体之间的相关性来进行实体嵌入，并通过简化来减少计算量。BDRI 提出了一种基于区块分解的关系交互模型，通过对知识图谱的二元张量表示进行区块项分解，增强了正反关系的融合，提高了模型的表达能力和计算效率。SAMI 模型将图注意力网络与张量分解结合起来学习知识图谱的表达，并通过自我对抗负采样技术加速训练，在链接预测任务中表现出色。

2.2 基于翻译模型的方法

基于翻译模型的方法是受到 Word2Vec 的启发，通过将实体和关系映射到低维向量空间中，从而将知识图谱中的关

系视为向量空间中的平移操作。

TransE 作为一个翻译模型, 将关系视为简单的翻译操作。虽然解决了大规模知识图谱的数据处理问题, 但其很难处理一些特定的关系, 如一对多、多对一、对称和传递等关系。TransH 通过将实体编码于特定于关系的超平面上, 在一对多、多对一关系上实现了更好的推理效果。此外, TransR 利用实体和关系的潜在空间, 在传递关系上取得了更为好的推理能力。TransD 首次利用大型知识图谱的实体和关系的独立投影向量来考虑可扩展性问题。图 4 给出了以上几种模型的示意图。

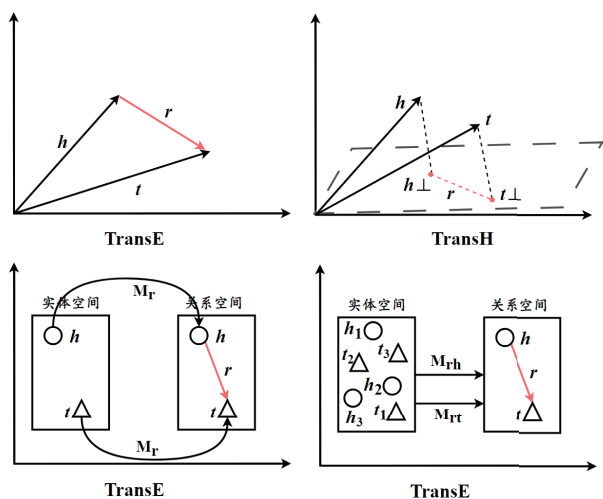


图 4 Trans 模型对比

还有一些方法尝试集成概率原理来建模知识图谱。例如, KG2E 利用高斯分布协方差, TransG 利用贝叶斯技术处理一对多关系问题。同时, TranSparse 为了缓解异构和不平衡问题, 通过设计自适应迁移稀疏矩阵来提高表达能力。MuRP 为实体嵌入投影设计了莫比乌斯矩阵向量乘法和莫比乌斯加法, 有效提高了精度和可扩展性。同时, TransW 利用词嵌入丰富了实体和关系嵌入, 在潜在的实体或关系推断中取得了很好的性能。RotatE 提出了一种基于旋转的复值嵌入的平移方法, 可以更好地推断对称性、反对称性等关系。此外, HAKE 对语义层次进行建模来代替基于极坐标空间的关系模式。TransRHS 首次将关系层次结构 (relational hierarchical structure, RHS) 整合到嵌入中。

此外, 为了使用统一的模型来处理更为复杂的关系事实, PairRE 用成对的向量对每个关系表示进行建模, 来对复杂关系进行自适应调整。HousE 设计了一种基于设计的豪斯霍尔德变换进行旋转和投影。还有一些针对更充分相互作用的翻译模型的尝试, 如 TripleRE 和 InterHT, 前者创造性的将关系向量分成 3 个部分, 利用残差的思想, 达到了更好的性能, 而后者通过增强尾实体和头实体之间的信息交互, 提高了模型的容量。MRE 模型提出了一种多关系嵌入的方法, 通过整

合直接关系、多跳关系和基于规则的关系来捕捉更丰富的语义信息。

3 知识图谱推理任务评价

本节主要介绍常用的知识图谱推理任务的数据集, 并介绍相关评价指标, 同时总结了基于分布式表示的知识图谱推理方法的实验结果。

3.1 数据集介绍

常用的知识图谱推理数据集按照来源分可以主要分成三类, 来自 Freebase、WordNet 和 YAGO。具体内容如表 1 所示。

Freebase 是一个庞大的开源知识图谱, 旨在构建一个全面的、结构化的知识库, 包含了关于各种实体及其关系的丰富信息。它的数据涵盖了从人物、地点到音乐、电影等多个领域, 支持复杂的查询和推理。Freebase 的数据主要来源于用户贡献和已存在的数据库整合, 对应数据集名称为 FB15k 和 FB15k-237。WordNet 是一个大型的英语词汇数据库, 其中的词被组织成同义词集, 每个同义词集表示一个独特的概念。WordNet 通过表示词与词之间的关系, 如同义词、反义词和从属关系, 提供了丰富的词汇信息, 广泛应用于自然语言处理和信息检索, 对应数据的名称为 WN18 和 WN18RR。YAGO 是一个基于 Wikipedia 和 WordNet 的知识图谱, 旨在整合这些来源的数据来构建一个高度结构化且精确的知识库。YAGO 提供了丰富的实体和关系信息, 广泛应用于语义网和知识管理领域, 对应数据集为 YAGO3-10。

表 1 常用数据集介绍

数据集名称	数据集来源	数据规模	构建方式	发表会议
FB15k	Freebase	1.4 万个实体 59 万条知识	从 Freebase 种去除掉反转了头实体和尾实体的关系而得到	NeurIPS 2013
FB15k-237	Freebase	1.4 万个实体 31 万条关系	在 FB15k 的基础上排除掉直接链接和冗余关系	EMNLP 2015
WN18	WordNet	4.1 万个实体 14 万条知识	从 WordNet 种挑取最常用的 18 种关系类型	NeurIPS 2013
WN18RR	WordNet	4 万个实体 9 万条知识	将 WN18 中的逆关系去除, 确保训练集和测试集中存在直接逆关系对	AAAI 2018
YAGO3-10	YAGO	12 万个实体 117 万条知识	是 YAGO3 的子集, 每个实体至少有 10 种关系	CIDR 2013

3.2 评价指标与实验结果

对于实体推理任务，模型的评价指标包括 MR、MRR 与 Hits@10 等。

MR（mean rank）通过计算测试集 T 中所有查询的平均排名来评估模型的推理性能，结果越接近 1（结果指标越大越好），说明效果越好，计算方法为：

$$MR = \frac{1}{|T|} \sum_{i=1}^{|T|} rank_i$$

(1)

此外，MRR（mean reciprocal ranking）通过计算测试集 T 中所有查询的平均倒数排名来评估模型性能，结果指标越大，说明效果越好，计算方法为：

$$MRR = \frac{1}{|T|} \sum_{i=1}^{|T|} \frac{1}{rank_i}$$

(2)

Hits@N（H@N）通过计算模型给出结果的在 Top N 中预测正确的查询数量占测试集 T 的比重来评估模型性能，结果指标越大，说明模型推理性能越好，计算方法为：

$$H@N = \frac{1}{|T|} \sum_{i=1}^{|T|} \mathbb{I}(rank_i \leq N)$$

(3)

式中： $\mathbb{I}(\cdot)$ 是 indicator 函数（若条件真则函数值为 1，否则为 0）。

表 2 总结了基于分布式表示的知识图谱推理方法在 WN18RR 和 FB15k-237 数据集上实验结果。

表 2 基于分布式表示的知识图谱推理方法实验结果总结

方法	WN18RR				FB15k-237			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
RESCAL	0.455	0.419	0.461	0.493	0.353	0.264	0.385	0.528
DisMult	0.420	0.370	0.439	0.521	0.243	0.191	0.271	0.328
ComplEx	0.440	0.410	0.460	0.510	0.346	0.256	0.386	0.525
TruckER	0.470	0.443	0.482	0.526	0.358	0.266	0.394	0.544
QuatE	0.481	0.436	0.500	0.564	0.311	0.221	0.342	0.495
DualE	0.482	0.440	0.500	0.561	0.330	0.237	0.363	0.518
HopfE	0.472	0.413	0.500	0.586	0.343	0.247	0.379	0.534
BDRI	0.495	0.456	0.513	0.563	0.370	0.276	0.406	0.553
SAMI	0.497	0.461	0.512	0.578	0.375	0.280	0.407	0.560
TransE	0.231	0.021	0.409	0.533	0.289	0.193	0.326	0.478
MuRP	0.481	0.440	0.495	0.566	0.335	0.243	0.367	0.518
RotatE	0.476	0.428	0.492	0.571	0.338	0.241	0.375	0.533
HAKE	0.497	0.453	0.515	0.582	0.335	0.237	0.371	0.530
PairRE	—	—	—	—	0.351	0.256	0.387	0.544
TripleRE	—	—	—	—	0.351	0.251	0.392	0.552
HousE	0.511	0.465	0.528	0.602	0.361	0.266	0.399	0.551
MRE	—	—	—	—	0.399	0.257	0.406	0.612

4 知识图谱推理在企业质量知识库中的应用

企业质量知识库是一个关于研制生产过程发生的质量问题的数据库，用于记录、存储、检索并共享信息。企业质量知识库是企业内、外部知识的集合，可以解决企业员工搜索有关的质量问题或疑问的答案，从而帮助员工更有效、有针对性地完成工作，把精力用于关键的任务和事情上。质量知识库包含了企业在质量管理方面所掌握的专业知识和实践经验，可以包括但不限于产品和服务质量标准、质量管理流程、质量控制技术、质量改进工具、质量文档和记录等。

在企业生产研发过程中，质量问题和投诉的记录往往是通过员工口头或书面形式传递的，存在信息流失和传递不及时的问题。由于企业人员流动性较大，相关的质量资料和经验也常常随之流失。对于企业来说，把质量管理全部寄希望于某个员工显然是不现实的。因此，需要一个系统化的管理，在系统中任何人都可以通过模糊搜索找出相关问题。

企业质量知识库与基于知识图谱推理的结合可以进一步提升其应用价值。通过将企业质量知识库中的数据与知识图谱推理技术相结合，可以实现以下几个方面的应用：

（1）智能搜索与问题诊断：利用知识图谱推理技术，员工可以通过自然语言查询快速获取质量问题的解决方案。例如，通过输入“某产品的常见质量问题”，系统可以基于历史数据和知识推理提供详细的解答和建议。

（2）知识共享和传承：通过知识图谱推理技术，可以将企业内外部的质量管理知识进行系统化地整理和分析，形成结构化的知识图谱，实现知识的共享和传承，避免因人员流动造成的知识流失。

（3）支持决策：管理层可以通过知识图谱推理获取全面的质量信息和数据支持，辅助决策，提升管理水平。例如，系统可以基于知识图谱推理提供关于某产品质量改进的详细报告，支持管理层的决策过程。

企业质量知识库与知识图谱推理的结合不仅是知识管理的重要工具，也是企业实现质量管理信息化和智能化的基础。未来，随着知识图谱推理技术的发展和运用，企业质量知识库将能够更高效地组织和利用企业内外部的质量管理知识，为企业的持续改进和创新提供有力支持。

5 总结与展望

本文对基于分布式表示的知识图谱推理技术研究进行了广泛的陈述，主要包括基于张量分解和基于翻译模型的方法。同时本文总结了每种类型方法的优缺点，如表 3 所示。

表 3 基于分布式表示的知识图谱推理方法优劣

推理方法	优势	不足
张量分解	表达能力强, 理论基础坚实, 灵活性强	计算复杂, 难以扩展
翻译模型	计算简单, 易于理解, 扩展性强	表达能力有限, 过于依赖低维嵌入

基于张量分解的方法, 如 RESCAL、DisMult、ComplEx 等, 通过将知识图谱表示为三阶张量, 并利用张量分解技术来进行推理。这类方法的优势在于其强大的表达能力和理论基础。张量分解模型可以捕获更复杂的关系和交互模式并且有着深厚的数学背景, 能够在理论上解释其表现。但同时张量分解模型在处理大规模数据集时计算复杂度较高, 并且难以在实际应用中扩展。基于翻译模型的方法, 如 TransE、TransH、TransR 等, 通过将实体和关系嵌入到低维向量空间中, 并将关系视为向量之间的平移操作。这种方法的优点在于其直观性和计算效率。翻译模型的计算复杂度较低, 适合大规模数据集, 且模型的直观性使其易于解释和实现。但由于翻译模型的简单性, 在处理复杂关系时存在局限, 且过度依赖低维嵌入, 可能会导致信息丢失, 难以捕捉复杂的关系模式。

现有的知识图谱推理方法在处理静态和单一模态数据方面虽已取得显著进展, 但在面对现实世界中更加复杂和多样化的信息需求时, 显得力不从心。未来的知识图谱推理研究将从多模态和时序两个方面探索和创新。

总的来说, 基于翻译模型和张量分解的知识图谱推理方法各有千秋, 未来的研究将逐步融合多模态和时序信息, 以提升推理的准确性和应用的广泛性。通过不断创新和融合, 知识图谱推理将更好地服务于各类智能应用, 为人工智能的发展提供坚实的基础。

参考文献:

[1] MONDAL I, HOU Y F, JOCHIM C. End-to-end construction of NLP knowledge graph[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. NewYork: ACM, 2021: 1885-1895.

[2] HUANG X, ZHANG J Y, LI D C, et al. Knowledge graph embedding based question answering[C]//Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. NewYork: ACM, 2019: 105-113.

[3] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD

international conference on Management of data.NewYork: ACM, 2008: 1247-1250.

[4] AUER S, BIZER C, KOBILAROV G, et al. DBpedia: a nucleus for a web of open data[J].The semantic web, 2007, 4825: 722-735.

[5] SUCHANEK F M, KASNECI G, WEIKUM G. YAGO: a large ontology from wikipedia and wordnet[J]. Journal of web semantics, 2008, 6(3): 203-217.

[6] CARLSON A, BETTERIDGE J, KISIEL B, et al. Toward an architecture for never-ending language learning[C]//Proceedings of the AAAI conference on artificial intelligence. NewYork: ACM, 2010: 1306-1313.

[7] SONG Q, WU Y H, DONG X L, et al. Mining summaries for knowledge graph search [C]//2016 IEEE 16th International Conference on Data Mining (ICDM). Piscataway: IEEE, 2016: 1215-1220.

[8] WANG X, HE X N, CAO Y X, et al. KGAT: knowledge graph attention network for recommendation[DB/OL]. (2019-06-08)[2024-07-12]. <https://doi.org/10.48550/arXiv.1905.07854>.

[9] 吴飞, 韩亚洪, 李玺, 等. 人工智能中的推理: 进展与挑战[J]. 中国科学基金, 2018, 32(3): 262-265.

[10] NICKEL M, TRESP V, KRIEGEL H P. A three-way model for collective learning on multi-relational data[C]//Proceedings of the 28th International Conference on International Conference on Machine Learning. Madison, WI, United States: ICML, 2011:809-816.

【作者简介】

陈光希(1979—), 女, 辽宁海城人, 硕士, 高级工程师, 研究方向: 质量管理、产品检验。

冯峰(1979—), 通信作者(email: ff_982002@sina.com), 女, 河南武陟人, 硕士, 正高级工程师, 研究方向: 网络安全、数据安全。

(收稿日期: 2024-11-18)