

知识图谱链路预测任务综述

许海涛

(华中科技大学, 计算机科学与技术学院, 武汉 中国 0085400)

摘 要 知识图谱 (Knowledge Graphs, KGs) 在工业界和学术界中有着广泛的应用, 这反过来又激发了从不同来源进行大规模信息提取的大量研究工作, 尽管做出了这些努力, 但众所周知, 即使是最大的知识图谱也存在不完备性问题; 链路预测 (Link Prediction, LP) 技术通过识别知识图谱中已有实体之间的缺失事实来解决这个问题。在最新的链路预测技术中, 基于知识图谱嵌入的链路预测技术在一些基准测试中取得了很好的效果。尽管关于这一主题的文献增长迅速, 但对于这些方法中设计选择的影响还没有引起足够的重视。此外, 这一领域的标准实践是通过聚合大量的测试事实来报告准确性, 在这些测试事实中, 一些实体比其他实体更具代表性; 这允许 LP 方法通过只关注此类实体的结构属性来显示良好的效果, 而往往忽略了 KG 的其余大部分。

本文的分析提供了基于嵌入的 LP 方法的综合比较, 扩展了分析的维度, 超过了文献中通常可用的范围。我们通过实验比较了 18 种最先进方法的有效性和效率, 考虑到基于规则的基线, 并报告了对于文献中最流行的基准的详细分析。

关键词 知识图谱; 链路预测; 知识图谱嵌入; 比较分析

中图法分类号 TP258

Overview of knowledge map link prediction tasks

Xu HaiTao

(Department of computer science and technology, HuaZhong University of science and technology, Wuhan, China)

Abstract Knowledge graphs (KGs) are widely used in industry and academia, which in turn has stimulated a lot of research work on large-scale information extraction from different sources. Despite these efforts, it is well known that even the largest knowledge graph has the problem of incompleteness; link prediction (link prediction, LP) technology solves this problem by identifying the missing facts between existing entities in the knowledge map. Among the latest link prediction technologies, the link prediction technology based on knowledge map embedding has achieved good results in some benchmarks. Although the literature on this subject is growing rapidly, it has no impact on the design selection of these methods. It has attracted enough attention. In addition, the standard practice in this field is to report accuracy by aggregating a large number of test facts, in which some entities are more representative than others; This allows LP methods to show good results by focusing only on the structural properties of such entities, often ignoring most of the rest of KG.

The analysis of this paper provides a comprehensive comparison of embedded LP methods, expands the dimension of analysis, and exceeds the scope commonly available in the literature. We experimentally compared the effectiveness and efficiency of 18 state-of-the-art methods, taking into account the rule-based baseline, and reported a detailed analysis of the most popular benchmarks in the literature.

Key words Knowledge atlas; Link prediction; Knowledge map embedding; comparative analysis

1 引言

知识图谱(Knowledge Graph, KG)是现实世界信息的结构化表示。在 KG 中, 节点代表实体, 如人和场所; 标签是可以连接它们的关系类型; 边是连接两个实体与关系的特定事实。由于它们能够以机器可读的方式对结构化数据建模, KG 被用于无数领域, 从问题回答到基于内容的推荐, 它们对任何语义 Web 项目都是至关重要的。值得注意的例子有 Freebase、Wikidata、DBpedia、Yago 以及业界的 Google KG、Microsoft Satori 和 Facebook Graph search。这些庞大的 KG 可以包含数百万个实体和数十亿个事实。

尽管有如此多的数据, 但众所周知, 即使是最富有的 KG 也存在不完整性的问题: 例如, 在 FreeBase 中, 超过 70% 的人实体没有已知的出生地, 超过 99% 的人没有已知的种族。这导致研究人员开始研究如何识别新的事实添加到 KG。这项任务通常被称为知识图谱补全或知识图谱增强, 可以通过从外部来源(如 Web 语料库)提取新事实或从 KG 中已有的事实推断来执行。后一种方法, 称为链接预测(Link Prediction, LP), 是我们分析的重点。

LP 任务一直是一个日益活跃的研究领域, 在近来它受益于机器学习和深度学习技术的爆炸性增长发展迅速。现在绝大多数的 LP 模型利用原始的 KG 元素来学习称为知识图嵌入的低维表示, 这些表示随后被用于推断新的事实。受 RESCAL 和 TRANSE 等少数开创性工作的启发, 在短短几年内, 研究人员已经开发了几十个基于非常不同的体系结构的模型。

这一领域的绝大多数论文有一个共同之处, 或者说存在一个共同问题, 那就是它们报告的结果汇总在大量测试事实上, 其中有少部分实体被过多地代表, 因此 LP 方法可以通过只关注这些实体而忽略其他实体来报告这些基准的良好结果。此外, 当前最佳实践的局限性可能会让人很难理解这篇文献中的论文是如何组合在一起的, 也很难描绘出哪些研究方向值得深入研究。换句话说, 当前技术的优点、缺点和局限性仍然是未知的, 而且几乎没有人调查过允许模型更好地执行的情况。粗略地

说, 我们仍然不知道是什么让一个事实变得容易或难以学习和预测。

为了缓解这些问题, 我们对一组具有代表性的基于 KG 嵌入的 LP 模型进行了广泛的比较分析, 特别关注仅利用 KG 结构来了解其嵌入的模型; 我们考虑 LP 任务种最先进的系统, 从无到有地训练和调整这些系统, 并通过提出新的、信息丰富的评估实践, 提供超出原始论文可用范围的实验结果。

我们考虑了属于不同体系结构的 18 个基于嵌入的模型, 并采用了一个额外的基于规则的最先进的系统作为基线。此外提供了实验比较方法的详细描述和相关文献的总结, 以及 KG 嵌入技术的分类。此外, 我们还考虑了五个最常用的数据集以及目前用于基准测试的最流行的指标; 我们详细分析了它们的特点和特性。对于每个模型, 我们提供每个数据集的效率和有效性的定量结果。我们在训练数据中定义一组结构特征, 并测量它们如何影响每个模型对每个测试事实的预测性能。

我们发现, 在所有数据集中, 模型对训练集中特定图形特征的存在有强烈的响应。它们的性能似乎对预测对等点特别敏感, 预测对等点是本工作中引入的一种新的图形结构, 它们对关系路径提供的逻辑支持特别敏感, 关系路径是我们使用原始度量估计的。此外, 我们还观察和研究了不同的模型在处理特定关系类型时如何表现出非常不同的行为。我们还发现, 在一些数据集中, 用于表示基数大于 2 的关系的过程在很大程度上改变了图的结构及其语义, 也影响了模型的性能。总之, 我们的研究结果表明, 对于 LP 模型, 哪些场景容易处理, 哪些场景仍需要进一步改进。我们最后发现, 由于低水平的方法差异, 一些模型的结果与其他模型没有直接可比性; 我们将解释这些策略如何影响评估过程, 并确定最负责任的体系结构组件。

2 链路预测问题

本节提供了关于知识图谱中的链路预测任务的详细概述, 介绍了我们在工作中提到的关键概念。我们将一个 KG 定义为一个标记的、有向的多图, 即多个不同标记的边可以连接相同节点的图: $KG = (E, R, G)$: 其中, E 表示实体的一组节点, R 表示关系的一组标签, $G \subseteq E \times R \times E$: 是一组边, 表示连接成对实体的事实。每个事实都是一个三元组 (h, r, t) , 其中 h 是头, r 是关系, t 是尾。LP 的任务

是利用 KG 中的现有事实来推断缺失的事实。这相当于在尾部预测中猜测 $\langle h, r, ? \rangle$ 缺失的正确实体,或在头部预测中猜测 $\langle ?, r, t \rangle$ 缺失的正确实体。更一般地说,在任何预测中,我们称之为已知实体为源实体,需要预测的实体为目标实体。随着时间的推移,已经研究了许多 LP 方法。有些方法基于可观察特征,依靠规则挖掘或路径排序算法等技术。随着机器学习的兴起,研究人员最近开始尝试通过学习图形元素的矢量化表示(称为 KG 嵌入)来捕捉图形的潜在特征。

通常,嵌入知识图谱中的是数字向量,可用于表示任何类型的元素(取决于域:单词、人、产品...)。它们是根据代表真实世界的数据集中相应元素的出现和相互作用的方式自动学习的。例如,基于单词在文本中的共现性,单词嵌入已经成为表示单词的标准方式。KG 嵌入用于表示 KG 中的实体和关系;它们体现了原始图的语义,因此可以用来识别其中的新链接。LP 数据集通常通过采样真实世界 KG 获得;因此,每个数据集都可以被视为一个小 KG,拥有自己的实体集 E 、关系 R 和事实 G 。为了便于研究, G 进一步被划分为训练集 G_{train} 、验证集 G_{valid} 和测试集 G_{test} 。所有基于嵌入的 LP 模型,为了估计事实的合理性,定义了一个计分函数 $\phi(h, r, t)$ 。在本文中,除非另有规定,否则我们将假设 ϕ 值越高,事实就越可信,缺失的实体往往被推断为得分最高的实体。

评估是通过通过对 G_{test} 中的所有测试三元组进行头尾预测,并在每次计算目标实体相对于其他实体的排名时进行的。排名可以在两种大不相同的设置下计算,分别是原始场景和过滤场景,这取决于处理排名高于目标答案的其他有效答案的方式。例如,当预测尾部时(巴拉克·奥巴马,父母,娜塔莎·奥巴马),模型可能会认为玛丽亚·奥巴马的得分高于娜塔莎·奥巴马,如图 1 所示。更一般地说,如果预测实体产生了包含在 G 中的事实,则它仍然可以被视为有效答案,即使它与目标答案不同。在原始场景中,超过目标实体的有效实体被视为错误,并且它们确实有助于排名计算。在过滤场景中,超过目标实体的有效实体不被视为错误,但是会在排名计算中跳过它们。

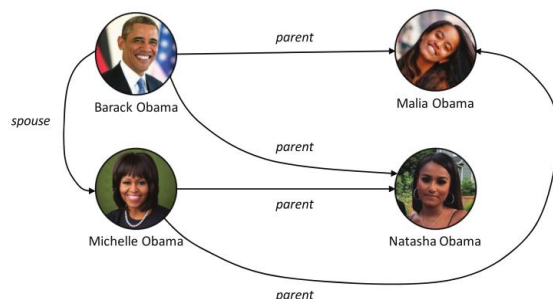


图 1 一个知识图谱中预测实例

3 链路预测技术综述

LP 任务可以依赖于多种方法,这取决于它们对优化问题的建模方式;为了概述它们高度多样的特征,我们提出了一种新的分类法,如图 2 所示,虚线箭头表示目标方法通过泛化或专门化其评分函数的定义,在源方法的基础上构建。我们定义了三个主要的模型族:矩阵张量分解模型族、几何方法模型族和深度学习模型族。它们中的每一个都被进一步分成更小的组,并以独特的颜色加以识别。对于每一组,我们在研究中包括最有效的代表性模型,优先考虑达到最先进性能的模型,并尽可能采用公开的实现。我们在对比分析的实验部分采用了这些模型。对于每个模型,我们还报告了它从其他模型受到的影响。我们相信,这种分类法有助于理解这些模型和我们工作中进行的实验。

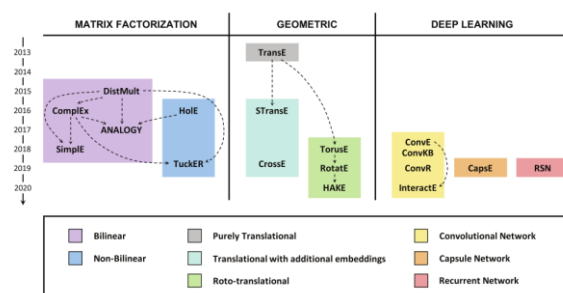


图 2 LP 任务模型分类

链路预测任务数据集通常通过对真实世界的知识图谱进行采样来获得,并将提取到的事实划分为训练集、验证集和测试集。主要在五个链路预测数据集上进行对比分析,表 1 是这五个数据集的比较分析及其一般性质。采用四种标准进行评判: MRR 、 MR 、 $H@1$ 和 $H@10$, MRR 是平均倒数排名, MR 是平均排名, $H@n$ 是链接预测中排名小于 n 的三元组的平均占比。

| | Entities | Relations | Triples | | | Reified Relations | Test Leakage | Multiple Domains |
|-----------|----------|-----------|---------|-------|-------|-------------------|--------------|------------------|
| | | | Train | Valid | Test | | | |
| FB15k | 14951 | 1345 | 483142 | 50000 | 50971 | X | X | X |
| WN18 | 40943 | 18 | 141442 | 5000 | 5000 | X | X | X |
| FB15k-237 | 14541 | 237 | 272115 | 17535 | 20466 | X | | X |
| WN18RR | 40943 | 11 | 86835 | 3034 | 3134 | | | |
| YAGO3-10 | 123182 | 37 | 1079040 | 5000 | 5000 | | | X |

表 1 比较分析中五个 LP 数据集一般性质

FB15k 数据集由 Bordes 等人生成,提取了所有 FreeBase 实体的 100 多个关系提及,并在 Wikilinks 数据库中显示,以及显示这些实体的所有事实。该数据集中排除了文字(如日期和专有名词),并将所有传递具体化 n 元关系的复合值类型(CVT)节点转换为二元边的因此操作极大地影响了图形结构和语义。

对于 WN18 数据集,它是从 WordNet4 中提取出来的,WordNet4 是一种语言 KG 本体,旨在提供一个字典/同义词表来支持 NLP 和自动文本分析。在 WordNet 中,实体对应于语法集(词义),关系表示它们的词汇连接。为了构建 WN18,使用 WordNet 作为起点,然后迭代地过滤掉很少提及的实体和关系。

FB15k-237 由 Toutanova 和 Chen 在发现 FB15k 由于存在等效和反向关系而遭受测试泄漏(即,模型在训练时看到的测试数据)后构建。他们评估了这个问题的严重性,展示了基于可观察特征的简单模型如何在 FB15k 上达到最先进的性能。然后,他们对 FB15k 进行了进一步采样,以创建一个更具挑战性的数据集,删除了所有等效或反向关系,以及提及最少的关系。为了过滤掉所有琐碎的三元组,他们还确保在训练集中连接的实体在验证集和测试集中没有直接链接。

WN18RR 是 Dettmers 等人在 WN18 试验中观察到类似于 FB15k 的泄漏后建立的 WN18 子集。他们证明了利用反向关系的简单反向模型在 WN18 和 FB15k 中都达到了最先进的结果,并通过从 WN18 中删除反向关系来构建 WN18RR。最近观察到,其测试集中的 212 个实体在培训中从未提及,因此约 6.7% 的测试事实无法合理预测。

YAGO3-10 是 Dettmers 等人提出的 YAGO3 样品。它是通过识别 KG 中至少有 10 种不同关系的所有实体,并提取所有相应的事实来构建的。正如作者所说,它的大多数三元组涉及人们的描述性特征,如公民身份或性别。YAGO3-10 上的反向模型性能较差。

3.1 张量分解模型

张量分解模型模型将 LP 解释为张量分解的任务。它们隐式地将 KG 视为一个三维(3D)邻接矩阵(一个三向张量),通过不完全性只能部分观察到。该张量可以分解为低维向量的组合,即实体和关系的嵌入。在实践中,嵌入只是通过利用已知事实进行事后学习,但它们也应该能够概括并将高分与看不见的真实事实联系起来。事实的分数可以通过对其元素的嵌入应用组合运算来计算。这些模型倾向于使用很少或根本没有共享参数;这使得它们特别轻便,易于训练。

在张量分解的双线性模型中,考虑到头部和尾部嵌入 $h \in \mathbb{R}^d$ 和 $t \in \mathbb{R}^d$,将关系嵌入表示为二维矩阵 $R \in \mathbb{R}^{d \times d}$ 。评分函数为双线性乘积: $\psi(h, R, t) = h^T R t$ 。(12) 这些模型不依赖于共享参数,并且通过在嵌入上引入特定的附加约束而彼此不同。我们参考以下具有代表性的模型:

1) DistMult, 该模型强制所有关系嵌入为对角矩阵,显著减少了参数空间。因此,计分函数是可交换的,其中 $\phi(h, R, t) = \phi(t, R, h)$: 这相当于将所有关系视为对称关系。尽管如此,当仔细调谐后,DistMult 仍能达到最先进的性能。

2) ComplEx, 该模型使用与 DistMult 相同的对角线约束,但在复数空间中扩展了其公式: $h \in \mathbb{C}^d$, $t \in \mathbb{C}^d$, $R \in \mathbb{C}^{d \times d}$, 因此,双线性乘积成为埃尔米特乘积,其中使用其共轭转置 t 来代替 t 。埃尔米特乘积是不可交换的,这使得 ComplEx 能够成功地模拟不对称关系。

3) Analogy, 该模型使用一般双线性公式,强约束模拟类比归纳结构: 其中 R 必须是正规矩阵: $RR^T = R^T R$; 且任何一对关系 R_1, R_2 的组合必须是可交换的: $R_1 \circ R_2 = R_2 \circ R_1$ 。作者证明,类比中的正规矩阵可以成功地学习非对称关系。

4) SimplE, 该模型使用与 DistMult 相同的对角线约束。它以正反两种形式对每个事实进行建模,平均它们的双线性乘积以计算总分。为了表示这种形式,它将每个实体 e 嵌入到单独的头向量和尾向量 e_h 和 e_t 中,并将每个关系 r 嵌入到单独的正向量和逆向量 R 和 R_{\perp} ,具有充分的表达能力,能够成功地模拟不对称关系;Lacroix 等人同时提出了一个类似的模型,称为 CP。

在非双线性模型中,往往使用不同于严格双线性积的公式将头、关系和尾嵌入结合起来。我们参考以下具有代表性的模型:

1) **HolE**, 该模型计算头部和尾部嵌入的循环相关性, 并将其乘以关系嵌入; 这可以看作是对全矩阵乘积的压缩, 使得 **HolE** 比无约束双线性模型更轻。已证明 **HolE** 在理论上等同于 **ComplEx**; 我们在工作中包括了这两种模型, 因为它们的不同公式会导致不同的时间复杂性, 并最终导致不同的结果。

2) **TuckER**, 该模型依赖于 **TuckER** 分解, 它处理独立维度 d_e 和 d_r 的实体和关系嵌入, 并共同学习共享核心 $W \in \mathbb{R}^{d_e \times d_r \times d_e}$ 。W 可以看作是原型关系矩阵的共享池。

| | FB15k | | | | WN18 | | | | FB15k-237 | | | | WN18RR | | | | YAGO3-10 | | | |
|-----------------------------|--------------|--------------|-----------|--------------|-------|-------|------------|-------|--------------|--------------|------------|--------------|--------------|--------------|-------------|-------|--------------|--------------|------------|--------------|
| | H01 | H010 | MR | MRR | H01 | H010 | MR | MRR | H01 | H010 | MR | MRR | H01 | H010 | MR | MRR | H01 | H010 | MR | MRR |
| Matrix Decomposition Models | | | | | | | | | | | | | | | | | | | | |
| DistMult | 73.61 | 86.32 | 173 | 0.784 | 72.60 | 94.61 | 675 | 0.824 | 22.44 | 49.01 | 199 | 0.313 | 39.66 | 50.22 | 5013 | 0.433 | 41.26 | 46.12 | 1107 | 0.501 |
| ComplEx | 82.32 | 91.04 | 33 | 0.855 | 94.43 | 96.15 | 190 | 0.951 | 27.18 | 55.80 | 144 | 0.387 | 44.77 | 58.06 | 2867 | 0.489 | 50.09 | 71.29 | 730 | 0.577 |
| ANALOGY | 65.59 | 83.74 | 126 | 0.726 | 92.61 | 94.42 | 808 | 0.934 | 12.59 | 35.38 | 476 | 0.202 | 35.82 | 38.00 | 9096 | 0.366 | 19.21 | 45.45 | 2423 | 0.283 |
| SimplE | 66.13 | 83.63 | 138 | 0.726 | 93.25 | 94.58 | 759 | 0.938 | 10.03 | 34.35 | 651 | 0.179 | 38.27 | 42.65 | 8764 | 0.308 | 35.76 | 43.16 | 2849 | 0.453 |
| HolE | 75.85 | 86.78 | 211 | 0.800 | 93.11 | 94.94 | 650 | 0.938 | 21.37 | 47.64 | 186 | 0.303 | 40.28 | 48.79 | 8401 | 0.432 | 41.84 | 45.19 | 6480 | 0.502 |
| TuckER | 72.89 | 88.88 | 30 | 0.788 | 94.64 | 95.80 | 510 | 0.952 | 25.90 | 53.61 | 162 | 0.352 | 42.95 | 54.40 | 6229 | 0.459 | 46.56 | 68.09 | 2417 | 0.544 |

图 3 张量分解模型指标效果

张量分解模型在五个数据集中的指标效果如图 3, 张量分解模型显示了跨数据集最可靠的结果, 在考虑到的实现方式中, 这些系统中的大多数系统在我们分析的数据集中的所有评估指标上均显示出统一的性能。特别是, **ComplEx** 及其 N3 正则化在所有数据集中的所有指标上显示了惊人的结果, 是唯一一个始终可与基线 **AnyBURL** 相比的基于嵌入的模型。

3.2 几何方法模型

几何方法模型将关系解释为潜在空间中的几何运算。头部嵌入根据关系嵌入的值进行空间变换 τ 。事实分数是结果向量和尾向量之间的距离, 使用距离函数 δ (例如, L1 或 L2 范数) 计算。根据 τ , 几何模型的分析形式可能类似于张量分解, 然而, 在这些情况下, 通常需要额外的约束使 τ 实现空间变换; 例如, 旋转操作的旋转可以表示为矩阵乘积, 但旋转矩阵必须是对角矩阵, 并且包含模数为 1 的元素。与张量分解非常相似, 几何模型倾向于避免共享参数, 直接在嵌入上运行反向传播。我们确定了几何方法族中的三个群体: 纯平移模型、具有附加嵌入的平移模型以及旋转平移模型。

纯平移模型: 这些模型将每个关系解释为潜在空间中的一个转换: 关系嵌入只是添加到头部嵌入中, 我们期望在靠近尾部嵌入的位置得到最终结果。因此, 这些模型将实体和关系表示为相同长度的向量。其中的 **TransE** 是第一个几何 LP 模型, 其灵感来自 **Word2vec** 嵌入的平移特性。它明确地强

制尾部向量接近头部和关系向量之和, 用 L1 或 L2 范数测量距离, 但是 **TransE** 无法正确模拟一对多、多对一、对称或传递关系。

具有附加嵌入的平移模型: 这些模型为每个 **KG** 元素学习多个嵌入; 他们通常为每个实体使用特定于关系的嵌入, 反之亦然, 为每个关系使用特定于实体的嵌入。这种附加嵌入的使用解决了纯平移方法的一些问题, 但代价是增加了参数的数量; 正如 **Kazemi** 和 **Poole** 所指出的, 它并不意味着完全的表达能力。其中的 **STransE** 模型, 除了在 **TransE** 中看到的 d 维嵌入外, 还将每个关系 r 映射到两个额外的 $d \times d$ 独立矩阵 W^h_r 和 W^t_r 。当计算事实的分数时, 在通常的转换之前, h 通过 W^h_r 和 t 乘以 W^t_r 计算所得; 这相当于对头部和尾部使用特定于关系的嵌入, 缓解了一对多和多对一关系的问题; 而 **Cross** 模型则为每个关系 r 学习相同大小的额外嵌入 c_r 。Cross 与深度学习模型和 **TuckER** 类似, 它使用非线性激活函数插入操作, 如双曲正切和 sigmoid。

旋转平移模型: 旋转平移模型包括不能直接表示为纯平移的操作: 这通常相当于以组合或代替平移的方式执行类似旋转的变换。其中的 **TorusE** 的动机是观察到 **TransE** 中的正则化, 迫使实体嵌入位于超球面上, 限制了它们满足平移约束的能力。为了解决这个问题, **TorusE** 将原始空间 $\mathbb{R}^{d_{in}}$ 的每个点 x 投影到环面 T^d 上的点 $[x]$ 。作者定义了分别对应于 L1、L2 和平方 L2 范数的圆环体距离 d_{L1} 、 d_{L2} 和 d_{L2^2} 。而 **RotatE** 将关系建模为复杂潜在空间中的旋转, 其中 h 、 r 和 t 都属于 \mathbb{C}^d 。嵌入的是一个旋转矢量: 在其所有元素中, 相位沿该轴传递旋转, 模数等于 1。 h 由 r 通过元素乘积旋转, 与 t 的距离用 L1 范数测量。旋转允许正确地建模许多关系模式, 例如对称/反对称、反转和合成。**HAKES** 则是通过将模数缩放与旋转相结合的关系来扩展旋转。**HAKES** 将每个 **KG** 元素映射为单独的模量和相位矢量。当在给事实打分时, 模块 h_m 通过 r_m 的点积进行缩放。与 t_m 和 t_p 的距离分别为 L1 范数和 \sin 。在分层关系中, 模数缩放占主导地位, 而非分层关系仅因其相位不同而不同。

| | | | | | | | | | | | | | | | | | | | | | |
|------------------|---------|-------|-------|-----|-------|-------|-------|-----|-------|-------|-------|-----|-------|-------|-------|------|-------|-------|-------|-------|-------|
| Geometric Models | TransE | 49.36 | 84.73 | 45 | 0.628 | 40.56 | 94.87 | 279 | 0.646 | 21.72 | 49.05 | 209 | 0.33 | 2.79 | 49.52 | 3036 | 0.266 | 40.57 | 67.39 | 1187 | 0.501 |
| | STransE | 39.77 | 79.60 | 69 | 0.545 | 43.12 | 95.45 | 208 | 0.656 | 22.48 | 49.56 | 337 | 0.315 | 10.13 | 42.31 | 5172 | 0.226 | 3.28 | 7.35 | 5797 | 0.040 |
| | CrossE | 60.08 | 86.23 | 136 | 0.780 | 73.26 | 95.01 | 441 | 0.834 | 21.21 | 47.05 | 227 | 0.298 | 38.07 | 44.99 | 5212 | 0.405 | 33.08 | 65.45 | 3839 | 0.446 |
| | TorusE | 68.85 | 83.98 | 143 | 0.746 | 94.33 | 95.44 | 525 | 0.947 | 19.62 | 44.71 | 211 | 0.281 | 42.68 | 53.35 | 4872 | 0.463 | 27.43 | 47.44 | 19455 | 0.342 |
| | RotatE | 73.93 | 88.10 | 42 | 0.780 | 94.30 | 96.02 | 274 | 0.940 | 23.83 | 53.06 | 178 | 0.336 | 42.60 | 57.35 | 3218 | 0.475 | 40.52 | 67.07 | 1830 | 0.498 |
| | HAKES | 74.49 | 88.44 | 43 | 0.786 | 94.34 | 96.08 | 280 | 0.950 | 24.01 | 54.28 | 184 | 0.347 | 45.28 | 58.08 | 3304 | 0.487 | 40.57 | 69.54 | 2068 | 0.546 |

图 4 几何方法模型指标效果

几何方法模型的效果如图 4 所示, 研究发现几何方法模型的性能稍微不稳定, 在过去的几年中, 研究投入了大量精力用于转换模型, 从 TransE 到它的许多具有多嵌入策略的后继者, 这些策略用于处理多对一, 一对多和多对多关系。这些模型显示出有趣的结果, 但仍然存在度量和数据集之间的一些不规则之处。例如, 像 TransE 和 STransEseem 这样的模型在 WN18RR 数据集上格外挣扎, 尤其是涉及 H@1 和 MRR 度量标准时。总而言之, 仅依靠翻译的模型已被最近的旋转翻译模型所取代。在这方面, RotatE 在所有数据集上都表现出非常一致的性能, 并且在考虑到 H@10 时尤为突出。

3.3 深度学习模型

深度学习模型使用神经网络来执行 LP 任务。神经网络学习到诸如权重和偏差等参数, 并将它们和输入数据相结合来识别重要模式。深度网络将参数组织到单独的层中, 通常穿插着非线性激活函数。随着时间的推移, 许多类型的层已经被开发出来; 例如, 密集层只需将输入数据 X 与权重 W 结合起来, 并添加一个偏差 $b: W \times X + b$ 。为了简单起见, 在下面的公式中, 我们将不提及偏差的使用, 使其保持隐式。更高级的层执行更复杂的操作, 如卷积层, 学习卷积核以应用于输入数据, 或以递归方式处理顺序输入的递归层。在 LP 领域, KG 嵌入通常与层的权重和偏差一起学习; 这些共享参数使这些模型更具表现力, 但训练时间可能更长, 更容易过度拟合。根据采用的神经体系结构, 在这个家族中确定了三组: 卷积神经网络、胶囊神经网络和递归神经网络。

卷积神经网络: 卷积神经网络模型依靠卷积层对输入数据 (例如, KG 元素嵌入到训练事实中) 进行卷积, 应用低维滤波器 ω 。结果是一个特征图, 然后通常将其传递到其他密集层以计算事实得分。ConvE 模型将实体和关系表示为一维大小的向量。在计算事实得分时, 它将 h 和 r 连接并重新整形为唯一的输入 $[h; r]$, 尺寸为 $d_m \times d_n$ 。这一输入通过一个卷积层和一组 $m\omega \times n\omega$ 的滤波器, 然后通过一个密集层和一组加权滤波器。最后使用点积将输出与尾部嵌入 t 结合, 得到事实分数。ConvKB 还将实体和关系建模为大小相同的向量; 它将 h 、 r 和 t 连接成一个 $d \times 3$ 矩阵 $[h; r; t]$ 。该输入通过一组 1×3 形状的滤波器进行卷积, 得到 $T \times 3$ 的特征图。特征图通过一个神经元和权重的密集层, 从而得到事实得分。ConvR 学习讲实体和关系嵌入作为

不同维度的向量来学习。InteractE 扩展并改进了 ConvE, Interactive 将重塑后的矩阵堆叠成 3D 张量, 然后使用深度循环卷积进行处理。

胶囊神经网络: 胶囊网络 (CapsNets) 由称为胶囊的神经元组成, 这些神经元组编码输入特定特征, 如, 图像中特定对象的存在。胶囊网络可以像卷积网络那样识别这些特征, 而不会丢失空间信息。每个胶囊将其输出发送到高阶胶囊, 连接由动态路由过程决定。CapsE 模型将实体和关系嵌入到相同大小的向量中, 并假设不同的嵌入在相同位置编码同源方面。它依赖于与 ConvKB 类似的体系结构, 但将最终的密集层替换为胶囊层, 在胶囊层中传递关于输入事实的一个方面的信息。带有一个胶囊的第二层用于产生三重评分。

递归神经网络: 递归神经网络 (Recurrent Neural Network, RNN) 模型采用一个或多个递归层来分析从训练集中提取的整个路径 (事实序列), 而不仅仅是单独处理单个事实。循环跳过网络 (RSN) 是专为 LP 设计的一种 RNN。由于实体和关系的变化, 基本 RNN 可能不适合 LP, 因此提出了一个新的层来积极考虑这一方面。由此产生的 RSN 模型通过有偏随机游动学习从训练集中提取整个路径, 并使用特别优化的基于类型的 NCE 计算损失。

| | | | | | | | | | | | | | | | | | | | | | |
|----------------------|-----------|-------|-------|-----|-------|-------|-------|-----|-------|-------|-------|-----|-------|-------|-------|------|-------|-------|-------|-------|-------|
| Deep Learning Models | ConvE | 59.46 | 84.94 | 51 | 0.688 | 93.89 | 95.66 | 413 | 0.945 | 21.90 | 47.62 | 281 | 0.305 | 38.39 | 50.75 | 4944 | 0.427 | 39.93 | 65.75 | 2429 | 0.498 |
| | ConvKB | 11.44 | 40.83 | 324 | 0.211 | 52.89 | 94.89 | 202 | 0.709 | 13.98 | 41.40 | 309 | 0.330 | 5.63 | 52.50 | 3429 | 0.249 | 32.16 | 60.47 | 1083 | 0.430 |
| | ConvR | 70.57 | 88.55 | 70 | 0.772 | 94.56 | 95.85 | 471 | 0.950 | 25.56 | 52.63 | 251 | 0.346 | 43.73 | 52.66 | 5646 | 0.467 | 44.62 | 67.33 | 2582 | 0.527 |
| | InteractE | 72.56 | 88.68 | 622 | 0.796 | 94.63 | 95.55 | 469 | 0.950 | 26.35 | 53.73 | 195 | 0.335 | 42.72 | 51.96 | 5056 | 0.459 | 46.61 | 68.51 | 2182 | 0.543 |
| | CapsE | 1.93 | 21.78 | 610 | 0.087 | 84.55 | 95.08 | 233 | 0.890 | 7.34 | 35.60 | 405 | 0.160 | 33.69 | 55.98 | 720 | 0.415 | 0.00 | 0.00 | 60674 | 0.000 |
| RSN | | 72.34 | 87.61 | 51 | 0.777 | 91.23 | 95.10 | 348 | 0.928 | 19.84 | 44.44 | 248 | 0.298 | 34.58 | 48.34 | 4210 | 0.395 | 42.65 | 66.43 | 1339 | 0.511 |

图 5 深度学习模型指标效果

深度学习模型在五个数据集上的指标效果如图 5 所示, 深度学习模型通常是最多样化的家族, 其结果因模型的架构选择以及它们的实现而异。迄今为止, ConvR 和 RSN 展示了该系列中最好的结果, 在 FB15k, WN18 和 YAGO3-10 中实现了非常相似的最新性能。在 FB15k-237 和 WN18RR 中, 其筛选过程已经切断了最相关的路径, RSN 似乎更困难, 这可能是由于其明确表达了利用路径的方式。另一方面, 诸如 ConvKB 和 CapsE 之类的模型在 H@10 和 MR 指标上取得了较好的结果, 但它们似乎在 H@1 和 MRR 上的表现不是很好。可以发现, 在 LP 任务中, 基于规则的 AnyBURL 被证明是性能出色的模型, 因为它的实验测试结果表明它始终处于最佳模型之列。

5 总结与展望

在知识图谱链路预测任务中，我们对于基于知识图谱潜入的 LP 模型进行了广泛的比较分析，在文献中五个最流行的数据集上分析了它们的效率和有效性，

在我们的分析中，张量分解模型显示了跨数据集最可靠的结果，在考虑到的实现方式中，这些系统中的大多数系统在我们分析的数据集中的所有评估指标上均显示出统一的性能。特别是，**ComplEx** 及其 **N3** 正则化在所有数据集中的所有指标上显示了惊人的结果，是唯一一个始终可与基线 **AnyBURL** 相比的基于嵌入的模型。另一方面，几何方向稍微不稳定，几何模型仍然存在度量和数据集之间的一些不规则之处。例如，像 **TransE** 和 **STransEseem** 这样的模型在 **WN18RR** 数据集上格外挣扎，尤其是涉及 **H@1** 和 **MRR** 度量标准时。总而言之，仅依靠翻译的模型已被最近的旋转翻译模型所取代。在这方面，**RotatE** 在所有数据集上都表现出非常一致的性能，并且在考虑到 **H@10** 时尤为突出。

深度学习模型的结果因模型的架构选择以及它们的实现而异。迄今为止，**ConvR** 和 **RSN** 展示了该系列中最好的结果，在 **FB15k**，**WN18** 和 **YAGO3-10** 中实现了非常相似的最新性能。在 **FB15k-237** 和 **WN18RR** 中，其筛选过程已经切断了最相关的路径，**RSN** 似乎更困难，这可能是由于其明确表达了利用路径的方式。另一方面，诸如 **ConvKB** 和 **CapsE** 之类的模型在 **H@10** 和 **MR** 指标上取得了较好的结果，但它们似乎在 **H@1** 和 **MRR** 上的表现不是很好。可以发现，在 LP 任务中，基于规则的 **AnyBURL** 被证明是性能出色的模型，因为它的实验测试结果表明它始终处于最佳模型之列。

参考文献

- [1] Yanjie Wang, Daniel Ruffinelli, Rainer Gemulla, Samuel Broscheit, and Christian Meilicke. 2019. On evaluating embedding models for knowledge base completion. In Proceedings of the 4th Workshop on Representation Learning for NLP.
- [2] Hung Nghiep Tran and Atsuhiko Takasu. 2019. Analyzing knowledge graph embedding methods from a multiembedding interaction perspective. In Proceedings of the EDBT/ICDT 2019 Joint

Conference.

- [3] Chandrhas, Aditya Sharma, and Partha P. Talukdar. 2018. Towards understanding the geometry of knowledge graph embeddings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.
- [4] Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating robustness and interpretability of link prediction via adversarial modifications. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.