知识图谱补全方法总结

知识图谱补全(knowledge graph completion)目的是预测出三元组中缺失的部分，从而使知识图谱变得更加完整.根据补全类型分类，知识图谱补全可以分为实体预测、关系预测、属性预测等.根据所需补全三元组中实体和关系是否均属于某一知识图谱，可以把知识图谱补全分成静态知识图谱补全和动态知识图谱补全.根据是否借助知识图谱之外的信息来补全缺失信息，分为内部方法和外部方法.根据补全方法分类，可分为基于逻辑符号、基于表示学习和基于图的知识图谱补全.

1 基于逻辑符号的知识图谱补全

基于符号逻辑的知识推理主要包括一阶谓词逻辑、描述逻辑和规则等.符号逻辑表示是最早使用的知识表示方法之一，且与人类的自然语言比较相近.基于符号逻辑的知识表示虽然可以很好地描述逻辑推理，然而计算机生成规则的能力很弱，往往需要大量的人力，而且对数据的质量要求较高.因此，在大数据时代，传统基于符号逻辑的方法在解决知识表示与推理问题时已经非常有限.

2 基于表示学习的知识图谱补全

目前对知识图谱的补全往往采用学习知识表示并定义得分函数对三元组进行打分的方法实现关系预测.这样一来，知识图谱的补全算法就成了对三元组的得分进行排序的算法.因此，介绍知识图谱补全算法亦即介绍知识表示模型算法.目前学习知识表示的代表模型有距离模型、神经网络模型、能量模型、双线性模型、张量神经网络模型、矩阵分解模型和翻译(转移)模型等.

2.1 距离模型

在结构表示(structured embedding，SE)中，每个实体用 d 维的向量表示，所有实体被投影到同一个 d 维向量空间中．同时，SE 还为每个关系 r 定义了 2 个矩阵，用于三元组中头实体和尾实体的投影操作．SE 将头实体向量和尾实体向量通过关系 r 的 2 个矩阵投影到 r 的对应空间中，然后在该空间中计算两投影向量的距离．这个距离反映了 2 个实体在关系 r 下的语义相关度，它们的距离越小，说明这 2 个实体存在这种关系．SE 模型有一个重要缺陷：它对头、尾实体使用 2 个不同的矩阵进行投影，协同性较差，往往无法精确刻画两实体与关系之间的语义联系.

2.2 神经网络模型

神经张量网络(neural tensor network，NTN)模型是是目前表示能力最强的嵌入表示模型.它的基本思想是，用双线性张量将不同维度下的头、尾实体的向量联系起来.由于 NTN 强大的表示能力，虽然能够更加精确地刻画实体及其关系之间的复杂语义关联，但是，由于计算复杂度非常高，所以并不适用于大规模的知识图谱.

单层神经网络模型(single layer model，SLM)是一种简化的神经网络模型，尝试采用单层神经网络的非线性操作，来减轻 SE 无法协同精确刻画实体与关系的语义联系的问题．虽然 SLM 是 SE 模型的改进版本，但是它的非线性操作仅提供了实体和关系之间比较微弱的联系，但却引入了更高的计算复杂度.

DSKG 模型提出了一种利用多层递归神经网络对知识图谱中的三元组进行序列建模的方法.由于模型基于序列特征，因此可在只有一个实体的情况下预测三元组.

SENN 模型是基于共享嵌入的 KGC 神经网络模型.它将头部实体、关系和尾部实体的预测任务集成到一个基于神经网络的框架中，共享实体和关系的嵌入，同时明确考虑这些预测任务之间的差异.提出了一种自适应加权损失机制，根据关系的映射特性和预测任务动态调整损失权重.由于关系预测通常比头尾实体预测性能更好，进一步将 SENN 扩展到 SENN+，利用它来辅助头尾实体预测.

2.3 能量模型

语义匹配能量模型(semanticmatchingenergy，SME)提出更复杂的操作，寻找实体和关系之间的语义联系.在SME 中，每个实体和关系都用低维向量表示．在此基础上，SME 定义若干投影矩阵，刻画实体与关系的内在联系．SME 为每个三元组(h,r,t)定义了 2 种评分函数，分别是线性形式和双线性形式此外，也有研究工作用三阶张量代替 SME 的双线性形式.

2.4 双线性模型

隐变量模型(latent factor model，LFM)提出利用基于关系的双线性变换，刻画实体和关系之间的二阶联系．通过简单有效的方法刻画了实体和关系的语义联系，协同性较好，计算复杂度低.

DISTMULT 模型还探索了 LFM 的简化形式：将关系矩阵设置为对角阵．实验表明，这种简化不仅极大降低了模型复杂度，模型效果反而得到显著提升.

2.5 翻译(转移)模型

TransE 模型是知识图谱补全算法中最为为经典的算法.

TransE 模型认为正确的三元组(h，r，t)(h 代表头实体的向量，r 代表关系的向量，t 代表尾实体的向量)需满足 h+r≈t，即尾实体是头实体通过关系平移(翻译)得到的.TransE 不适合对复杂关系进行建模，在复杂关系上的表现比较差.

TransH 模型在 TransE 的基础上为每个关系多学一个映射向量,用于将实体映射到关系指定的超平面;然后在该超平面,与 TransE 一样,关系表示向量看成映射后的实体之间的转移.映射向量使得对于不同关系,同一个实体在不同关系指定的超平面有不同的表示,一定程度上缓解了不能很好地处理多映射属性关系的问题.

TransR 模型认为实体和关系存在语义差异，它们应该在不同的语义空间.此外，不同的关系应该构成不同的语义空间，因此 TransR 通过关系投影矩阵，将实体空间转换到相应的关系空间.

CTransR 模型将关系划分为关系组,为每个关系组学习一个关系向量和映射矩阵.

TransD 模型认为头尾实体的属性通常有比较大的差异，因此它们应该拥有不同的关系投影矩阵.此外，考虑矩阵运算比较耗时，TransD 将矩阵乘法改成了向量乘法，从而提升了运算速度.

TranSparse(share)模型同样将关系看成实体之间的转移,但用自适应的稀疏转移矩阵替换一般的转移矩阵.转移矩阵的稀疏度由关系连接的实体对数目决定,头尾实体共享相同的转移矩阵.复杂关系的转移矩阵比简单关系更稠密，用来克服关系之间的异质性;TranSparse(separate)模型每个关系有两个单独的稀疏转移矩阵,头尾实体各一个,稀疏度由关系连接的头/尾实体数目决定，用来解决关系内部头尾实体的不均衡性问题.

TransAt 模型可以同时学习基于翻译的嵌入、与关系相关的实体类别和与关系相关的注意事项，用来解决以往模型只关注部分属性，忽略人类认知的层级规律问题.

由同一关系的三个一组中头部和尾部实体的性质不同，AEM 模型将每一个头部实体向量和每一个尾部实体向量分别由对应的头部关系向量和对应的尾部关系向量加权.然后得到了新的实体向量表示形式，且相同三元组中的新实体向量相似.由于 AEM 对实体向量的每个维度进行加权，因此能够准确地表示实体的潜在属性和关系.此外，AEM 的参数数量非常少，易于训练.

TransP 模型基于长短时记忆神经网络和已有的翻译模型，是一种多模块混合神经网络模型.TransP 通过对实体路径及其关系路径进行建模，可以有效挖掘实体间的间接关系，从而提高知识图完成任务的质量.

ProjR 模型将 TransR 和 ProjE 结合在一起，通过为每个关系定义一个唯一的组合算子来实现不同的表示.在 ProjR 中，具有不同关系的输入头实体-关系对将经历不同的组合过程.

2.6 基于张量/矩阵分解的模型

基于张量/矩阵分解的表示推理将(头实体,关系,尾实体)三元组看成张量/矩阵中的元素构建张量/矩阵,通过张量/矩阵分解方法进行表示学习.分解得到的向量表示相乘重构成张量/矩阵,元素值即为对应三元组有效与

否的得分,可以认为得分大于特定阈值的三元组有效,或候选预测按照得分排序,选择得分高的候选作为推理结果基于空间分布的补全.

RESCAL 模型基于三阶张量进行表示学习，模型虽然推理准确率高,但内存占用量大,计算速度慢.

TRESCAL 模型是在 RESCAL 的基础上引入实体类型信息这一关系域知识,在损失函数的计算中排除不满足关系特定的实体类型约束的三元组,加速计算.

ARE(additive relational effects)模型学习知识图谱三元组的隐性和观察到的模式,用一个附加项增广 RESCAL 模型对应观察到的模式.这里,观察到的模式是指用可观察的关系学习方法,例如规则方法等,得到的预测结果构成的三阶张量.附加项为该三阶张量乘以一个权重向量,权重向量衡量关系学习方法对各个关系的预测能力.该附加项通过减少不连接部分,降低 RESCAL 分解需要的阶.

TuckER 提出了一种基于知识图三元组二元张量表示的 TuckER 分解的相对简单但功能强大的线性模型.TuckER 在标准链接预测数据集上的性能优于所有以前的先进模型.TuckER 是一个完全表达模型，推导出了其实体上的边界和关系的嵌入维数，从而实现了完全表达，这比以往最先进的复杂和简单模型的边界小几个数量级.

2.7 空间分布模型

基于空间分布的表示推理建立模型拟合知识图谱中实体和关系的空间分布特征,使得在向量表示空间中,实体和关系的空间分布尽可能地与原知识图谱一致.该类方法通过设计对应的能够反映空间分布特征的得分函数,与简单地采用基于转移假设得分函数的基于转移方法区别开来,但采用与基于转移方法类似的学习和推理过程.

TransG 模型自动发现关系的隐性语义,利用关系不同隐性语义向量的混合转移头尾实体对,建模三元组,对应的产生式过程为尾头实体向量之差符合以每个关系隐性语义向量为均值、单位阵为协方差的高斯分布的加权和.TransG 可以学习出关系在对应三元组中最主要的隐性语义向量,指数作用拉大了其与关系的其他语义向量对三元组的贡献差距.

KG2E 是另一种高斯模型，转向基于密度的表示,直接建模实体和关系的确定性,在多维高斯分布的空间中学习知识图谱的表示.

2.8 引入实体文本描述的模型

DKRL 模型提出在嵌入模型中考虑知识图谱中提供的实体文本描述信息，即将每个实体对应的实体描述作为输入，并提出了 2 种模型：一种是利用 CBOW 将文本中的词向量做简单的相加操作作为文本表示，一种是利用卷积神经网络考虑了文本中词语的顺序信息.

STKRL 模型首先从语料库中收集提取每个实体相关的所有句子，通过词向量和神经网络模型对每个句子编码，最后学习得到每个实体基于序列文本的嵌入向量表示.与实体原有的嵌入向量表示一起构成了实体的向量表示.

Joint Embedding 模型利用远程监督将已知的三元组对齐到语料库当中，进而得到关系的文本路径.最后利用词向量和卷积神经网络对路径进行编码，得到关系路径的向量表示，最终参与到判别三元组的计算中.

TEKE 模型通过从语料库中收集与某个实体相关的信息，不仅构造了基于文本的实体嵌入表示，还构造了实体对的嵌入表示.最后所有的嵌入表示都参与到判别三元组的计算中.

CACL 模型直接利用了每个多跳邻居中包含的连接模式，因此可以更好地捕捉实体之间的结构角色相似性，从而获得更多的信息实体和关系嵌入.CACL 根据多跳邻居的相对重要性收集实体和关系作为上下文信息，并将它们唯一地映射到线性向量空间.卷积架构利用深度学习技术来表示每个实体及其线性映射的上下文信息.因此，可以从上下文中精心提取关键连通性模式的特征，并将它们合并到一个评估事实有效性的得分函数中.

ConMask 模型学习实体名称的嵌入及其文本描述的部分，以便将不可见的实体连接到知识图谱.为了减少

噪声文本描述的存在，ConMask 使用一个依赖关系的内容掩蔽来提取相关的片段，然后训练一个完全卷积的神经网络来将提取的片段与知识图谱中的实体融合在一起.

2.9 引入路径描述的模型

Path-Constraint Random Walk 模型和 Path Ranking Algorithm 模型等算法通过考虑关系路径信息在关系链接预测任务中取得了显著效果.PTransE 模型和 RTransE 模型等提出考虑关系路径来提升嵌入表示模型在知识图谱补全中的推理性能.

3 基于图的知识图谱补全

相比于嵌入表示学习模型，基于图的推理模型的优点在于解释性强，往往可以直接解释被预测三元组成立的原因.图特征模型一直在互联网链接关系预测、生物蛋白网络分析和社交网络分析等关系种类单一的链接预测任务中比较流行.这类方法认为每个结点的语义是由其周围结点确定的，且相似的结点更有可能存在关联，因此通过考虑邻居结点或结点间的路径信息更容易得出结点之间的语义关联.

GRank 模型为每个图模式构建了一个实体排名系统，并使用排名度量对它们进行评估.通过这样做，可以找到对预测事实有用的图形模式.

4 知识图谱通用数据集

数据集有 WordNet(WN11，WN18)，Freebase(FB13，FB15k，FB20k)，DBPedia (DBPedia50k，DBPedia500k)，LUBM，YAGO 等.