SEU 知识图谱表示学习

2019年11月30日 13:39

- **什么是知识表示学习**
  - 表示学习：将研究对象的语义信息表示为稠密低维的实值向量
    研究对象：文字（词汇、短语、句子、文章）、图片、语音等。
  - 知识表示学习：将知识库中的实体和关系表示为稠密低维的实值向量
  - 知识图谱包括实体和关系
    - 节点代表实体
    - 连边代表关系
  - 知识常用三元组进行表示
    - （head， relation， tail）
  - 计算效率问题
    基于图结构的知识表示虽然简洁直观，但是需要专门的图算法（复杂度高，可扩展性差）
  - 数据稀疏问题
    长尾分布，长尾上的实体和关系的语义难以捕获
  - 独热表示
    假设所有的研究对象都是独立的，将研究对象表示为向量，只要某一维是非零的，琦玉维度上的值均为0.显然不符合实际情况，导致丢失大量信息
    E.G.苹果（0,1,0,0,0,0,0） 香蕉（0,0,0,1,0,0,0）
- **知识表示学习的意义**
  - 低维向量提高计算效率
  - 稠密向量缓解数据稀疏
  - 多源的异质信息表示形式统一，便于迁移和融合

- **知识表示学习代表模型**
  - **基于距离的模型——UM**

- 仅利用头实体和尾实体的共现信息，而忽略了他们之间的关系。
  得分函数：

$$f(h,r,t) = -\| h - t \|_{l1/l2}$$

- **基于距离的模型——SE**
  - 与UM相比，SE添加了关系信息：将关系建模为分别针对头实体和为实体的矩阵。得分函数：

$$f(h,r,t) = -\| M_r^h h - M_r^t t \|_{l1/l2}$$

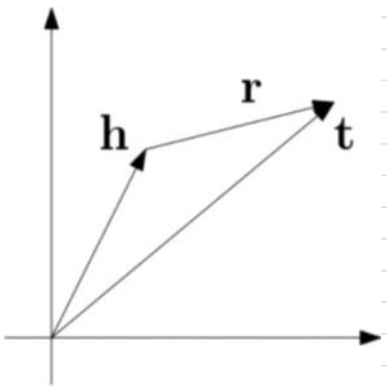
- **基于翻译的模型——TransE**
  - 对每条知识（head，relation，tail）中的relation看作从head到tail的翻译操作，得分函数：

$$f(h,r,t) = \| h + r - t \|_{l1/l2}$$

例子：

（北京，是…的首都，中国）

（华盛顿，是…的首都，美国）



$$L = \sum_{\xi \in T} \sum_{\xi' \in T'} [\gamma + f(\xi) - f(\xi')]_+$$

TransE存在的问题：不能很好的处理复杂关系，如1-N N-1 N-N关系

E.G.TransE会把“奥巴马” “布什”的向量变得相同

  - 存在问题还有：得分函数只采用了L1 L2距离，灵活度不够
  - 损失函数过于简单，实体和关系向量的每一维等同考虑

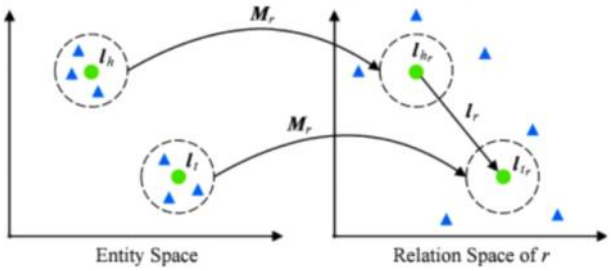
- **基于翻译的模型——TransH**
  - 为了解决TransE无法很好处理复杂关系的问题，TransH将每种关系建模为一个超平面，将三元组中的头实体和尾实体分别映射到该超平面中。得分函数：

$$h_\perp = h - w_r^T h w_r, \quad t_\perp = t - w_r^T t w_r$$

$$f(h,r,t) = -\| h_\perp + r - t_\perp \|_{l1/l2}$$

- **基于翻译的模型——TransR/CTransR**
  - TransE和TransH都将实体和关系嵌入到同一个向量空间中
  - TransR认为不同的关系矩阵应该有不同的语义空间。将期关系嵌入为矩阵，将头尾实体映射到子关空间中
  - CTransR将属于同一关系的头尾实体对分成聚成多个类，针对每个类学习不同的关系矩阵

$$h_\perp = hM_r, \quad t_\perp = tM_r$$



原来在实体空间中与头尾实体（圆圈）相似的实体（三角形），在关系r的子空间中被区分开了。

- 在TransH和TransR/CTransR中，不同的种类的实体共享相同的映射向量或矩阵，但一个关系的头尾实体的种类和属性往往有较大的差距
- TransR引入了控件映射，但是模型参数急剧增加
- TransD将每个对象（实体、关系）嵌入为两个向量：语义向量、映射向量

- **TranSpace**
  - 克服关系的异质性（有的关系链接的头尾实体对较多，有的较少）和不平衡性（同一关系的头实体和尾实体的数量不对称）
  - TranSpace用自适应的稀疏矩阵代替一般的映射矩阵，稀疏度由关系链接的头尾实体对的数量决定
- **TransM**
  - 放宽了前面几种模型所使用的基本条件，即：h+r≈t，方法是在前面得分函数的基础上加了权重：

$$w_r = \frac{1}{\log(h_r pt_r + t_r ph_r)}$$

- ManiFoldE
  - 将约束h+r≈t放宽为一种局域流形的约束。得分函数是：

$$f(h,r,t) = \| M(h,r,t) - D_r^2 \|_{l1/l2}$$

$$M(h,r,t) = \|\mathrm{h+r-t}\|_{l2}$$

- TransF
  - TransF将约束h+r≈t放宽为：只要求h+r（或t-r）的方向与t（或h）一致。得分函数同时衡量了h+r和t t-r和h的方向：

$$f(h,r,t) = (h+r)^T t + (t-r)^T h$$

- TransA
- KG2E
- TransG

- **语义匹配模型**
  - **LFM**
    - 基于关系双线性变换。协同性较好，计算复杂度低

$$f(h,r,t)=h^T M_r t$$

  - **DistMult**
    - 将LFM中关系的表示矩阵限制为对角阵，这种简化极大的降低了模型复杂度，模型效果范围得到提升。

$$f(h,r,t)=h^T diag(M_r)t$$

    - ANALOGY
    - RESCAL
    - HoIE
    - SLM
    - SME
    - NTN
    - MLP
    - NAM
    - ConvE（好多都看不懂……以后回来看）

- **实体类别**
  - SSE假设属于同一类别的实体在表示空间中应该距离较近，其使用流形学习算法对这种假设进行建模
  - TKRL利用层次类别信息建立实体的映射矩阵
- **文本描述**
  - **NTN**：先用辅助新闻语料库词汇进行表示，然后用实体中词汇的表示向量的平均值来初始化实体的表示。例如，用smart和phone的表示向量的平均值初始化实体smartphone的向量表示
  - **DKRL**：每个实体有两个表示，分别是基于结构的表示es和基于文本描述的表示ed。得分函数：

$$f(h,r,t)=-\|h_s+r-t_s\|-\|h_d+r-t_d\|-\|h_s+r-t_d\|-\|h_d+r-t_s\|$$

  - **TEKE**：将知识图谱中的实体与文本库中的词汇对齐，用在文本库中与头尾实体对应的词汇共现的词汇作为该三元组的文本上下文。
- **逻辑规则**
  - **ILP**将推理看做整数线性规划问题已解决知识图谱补全，目标优化函数包括两个部分：具体的KGE模型的优化函数，基于规则的一系列约束
  - **KALE**将三元组事实表示为原子公式，将规则表示为基于三元组事实的复合公式，然后用一个损失函数同时优化二者

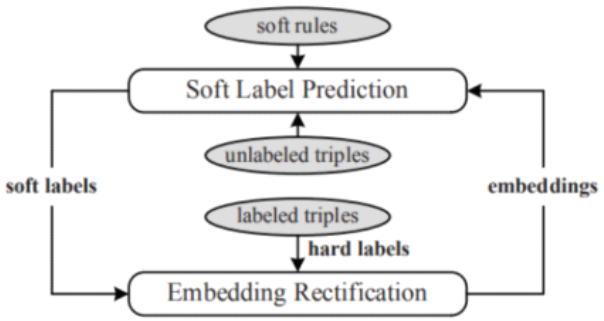
$$I(h,r,t)=1-\frac{1}{3\sqrt{d}}\|h+r-t\|$$

$$I(f_1 \Rightarrow f_2)=I(f_1)\cdot I(f_2)-I(f_1)+1$$

$$I(f_1 \wedge f_2 \Rightarrow f_3)=I(f_1)\cdot I(f_2)\cdot I(f_3)-I(f_1)\cdot I(f_2)+1$$

其中，$I(h,r,t) \in [0,1]$代表三元组为真的概率。

- **RUGE**重复迭代下述步骤完成对embedding的训练：
  - 由已知得分的三元组及规则，为未出现的三元组生成预测得分
  - 使用当前emdedding为每个三元组生成计算得分，并使用梯度下降法最小化计算得分与预测/实际得分的距离来对embedding进行更新



RUGE模型框架

- **实体属性**
  - 大部分KGE模型将实体属性也看做关系处理，但实际中，部分属性是不可能出现在关系中的
- **时序信息**
  - E.G.（霍金，出生于，英国牛津），（霍金，逝世于，英国剑桥）
    出生于和逝世于有严格的时间先后顺序
  - 有模型在三元组的基础上，添加时间维度，构成四元组（h,r,t,T），构建时序转换矩阵M捕捉关系r1和r2的先后关联
- **图结构**
  - **GAKE**将知识图谱看作有向图，定义了实体的三种上下文，帮助捕获三元组的语义
    - 邻居上下文：实体的所有出边及出边所连接的尾实体
    - 边上下文：实体的所有出边
    - 路径上下文：包含实体的关键路径
- **最新进展**
  - TransC将知识图谱中的实例和概念区别对待
    - 实体嵌入为向量、概念嵌入为球体。用点和球，球和球之间的相对位置关系对instanceOf和subClassOf两种关系建模，普通关系采用TransE模型
  - TransN
    - 利用实体在知识图谱中的邻居节点你，将实体和关系分别嵌入为两种向量

        - 语义向量：用于表示实体或关系的语义
        - 上下文向量：用于表示其他实体或关系的上下文。TransN选择邻居节点（实体关系对）
        - 利用实体的邻居总数动态计算所选邻居的数量
        - 利用关系间的相似度动态计算邻居节点的权重
    - GAN-based Framework
        - 利用生成对抗网络中的生成器进行高质量的负采样，用鉴别器计算reward和进行表示学习。因而该框架可以应用于不同的传统模型。

- **模型评测**
    - **常用数据集-WordNet**
      WordNet是最著名的词典知识库，拥有极高准确率的本体知识，主要是用于词义消歧。其主要定义了名词，动词，形容词和副词之间的语义关系。
    - **Freebase**
      Freebase将WordNet与Wikipedia二者的知识相结合。Freebase的分类系统包含topic、Type、Property、Schema
    - **YAGO**
      综合型知识库，整合了Wikipedia、WordNet以及GeoNames等数据源
- 评测任务
    - **链接预测**
        - MRR：所有正确实力排名的倒数的平均值
        - Hits@N：正确实例的排名中不大于N的比例
    - **三元组分类**
        - 知识图谱中的三元组表示一个二分类问题
          判断给定的三元组是否是知识图谱中真实的存在
        - 评测标准通常使用评测标准使用准确率、精确率、召回率和F1值
    - **知识融合**
    - **人机交互**

- **知识表示学习的挑战**
    - 大规模知识图谱的在线学习：
      知识图谱动态演化速度之快，如何开展在线学习以及知识的分布式表示？
    - 融合知识图谱丰富信息的表示学习：
      目前已存在的模型大多只利用了知识图谱中的一部分知识，如何融合知识图谱中多模态异构数据来更

- 大规模知识图谱的在线学习：
  知识图谱动态演化速度之快，如何开展在线学习以及知识的分布式表示？

- 融合知识图谱丰富信息的表示学习：
  目前已存在的模型大多只利用了知识图谱中的一部分知识，如何融合知识图谱中多模态异构数据来更好地进行知识表示？