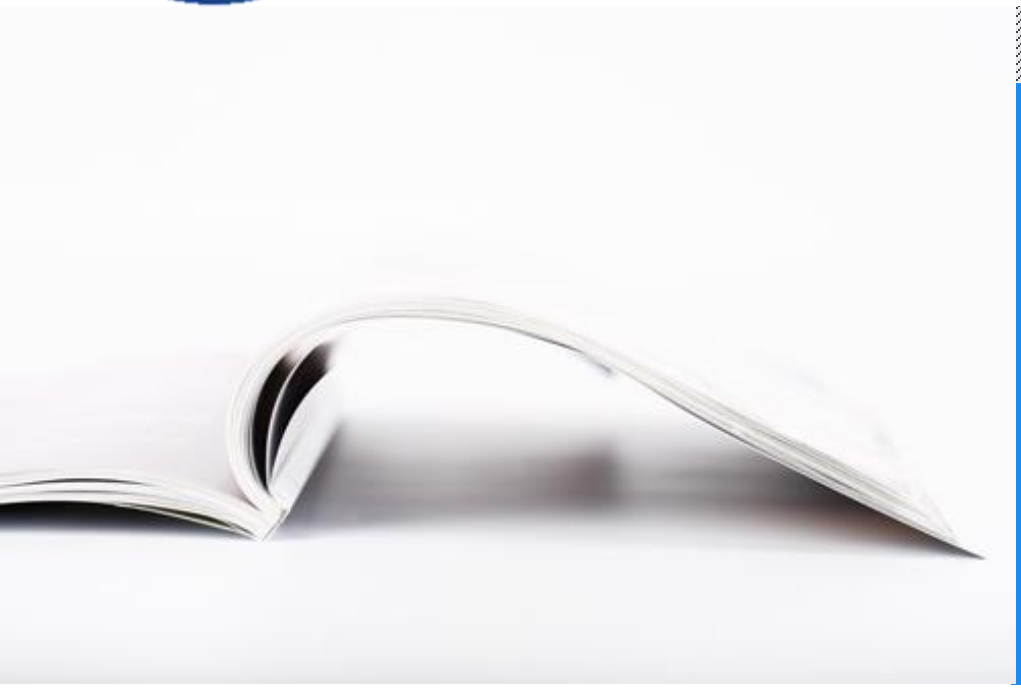


Web信息处理与应用

第十四节

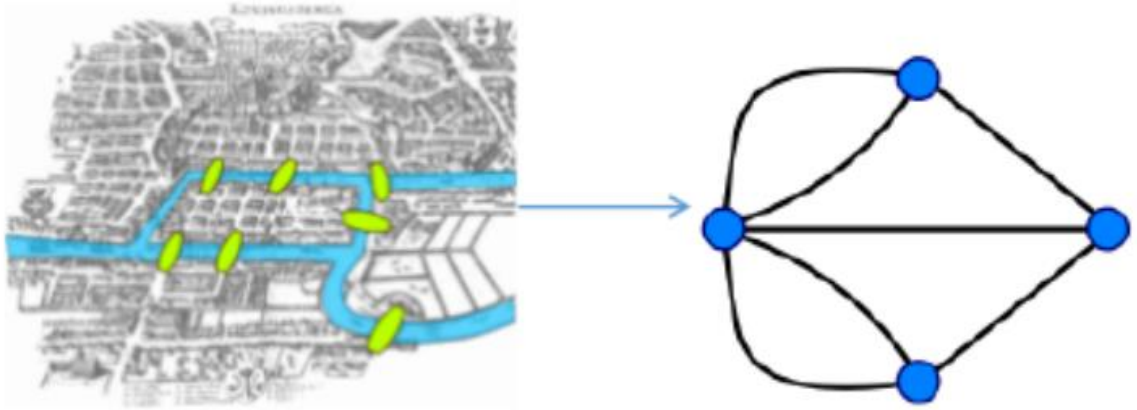
知识图谱与图计算

徐童

2024.12.12

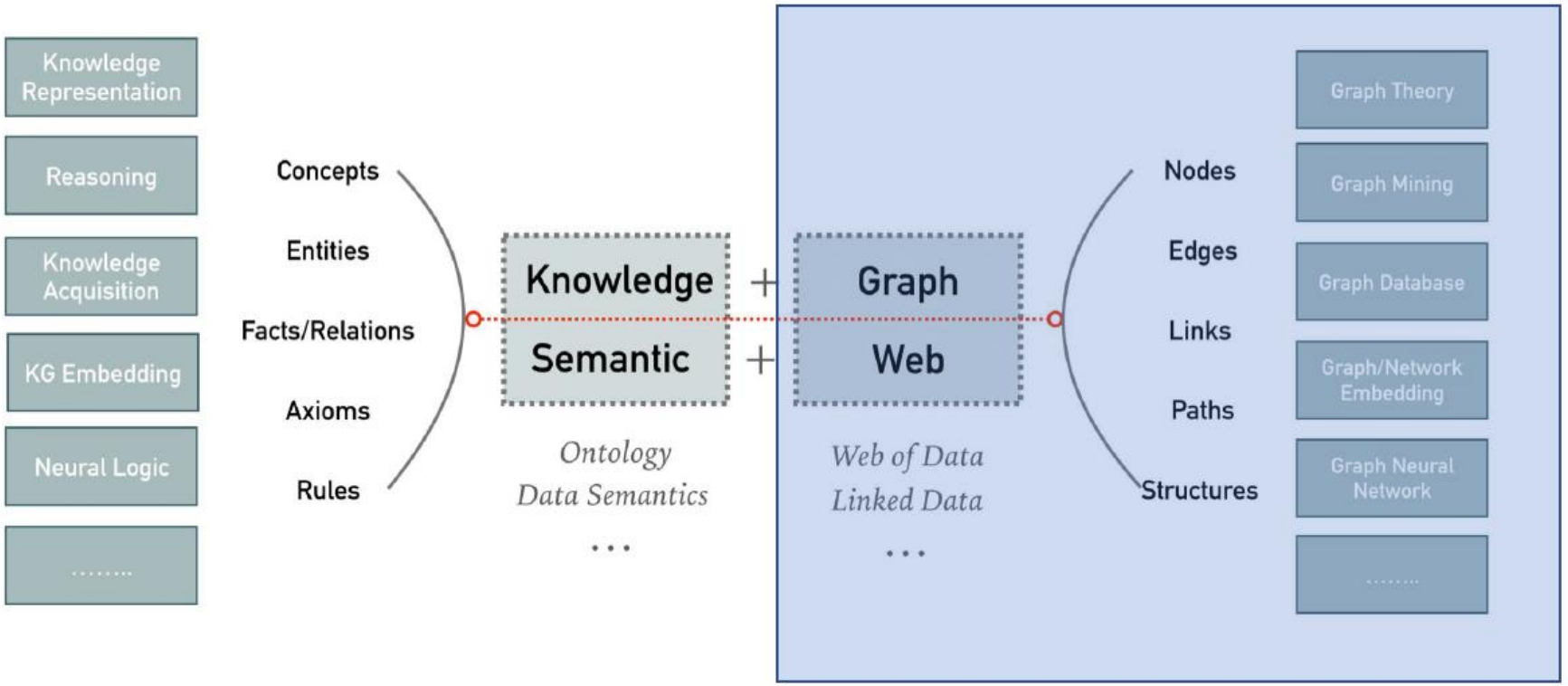
- **知识图谱的基本元素：边**

- 一般而言，知识图谱中的边用来表示关系（Relation）和属性（Attribute）。
  - 关系：侧重实体（Entity）之间的关联，例如“高于”：姚明高于小四
  - 属性：用于描述实体的特征，例如尺寸，颜色、组成等等
- 点和边组成知识图谱的基本单位：**三元组**（实体-关系-实体）



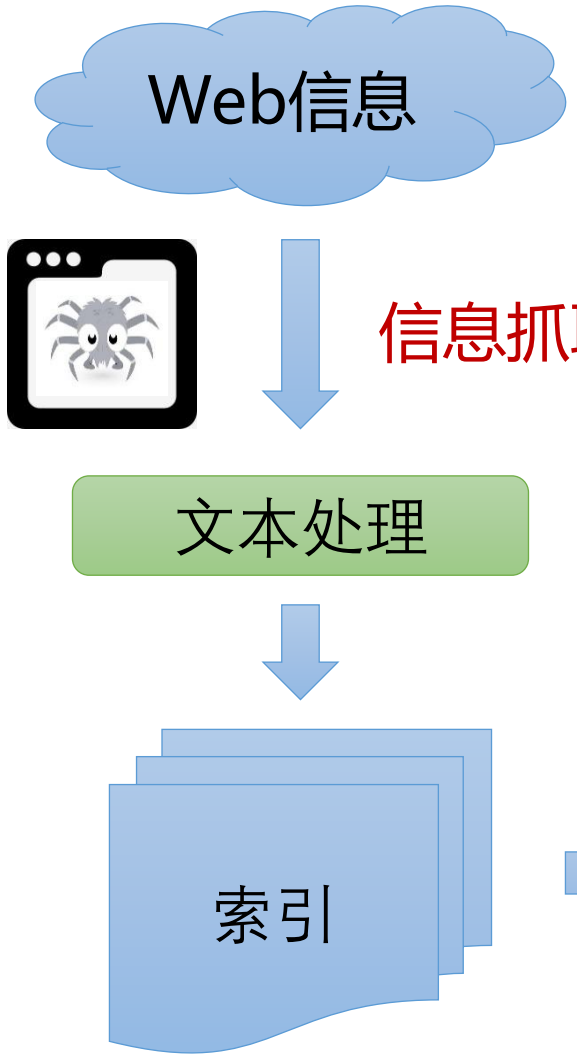
- **知识+图谱：两个维度的技术互补**

- 与传统的图结构相比，知识图谱具有更强的表现力，但结构上却没有超出图数据的范畴，仍然可以将图相关计算适用到知识图谱的研究中来



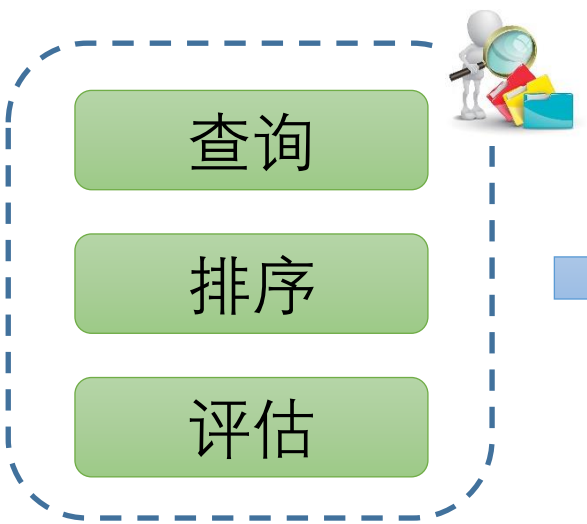
本节主题

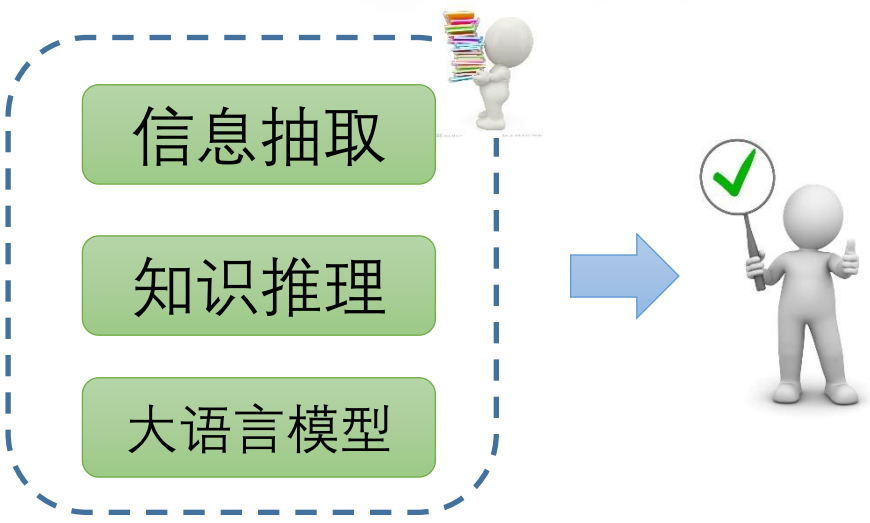
- 本课程所要解决的问题



第十三个问题：

如何将知识图谱与图学习技术相结合，支撑图谱推理应用？



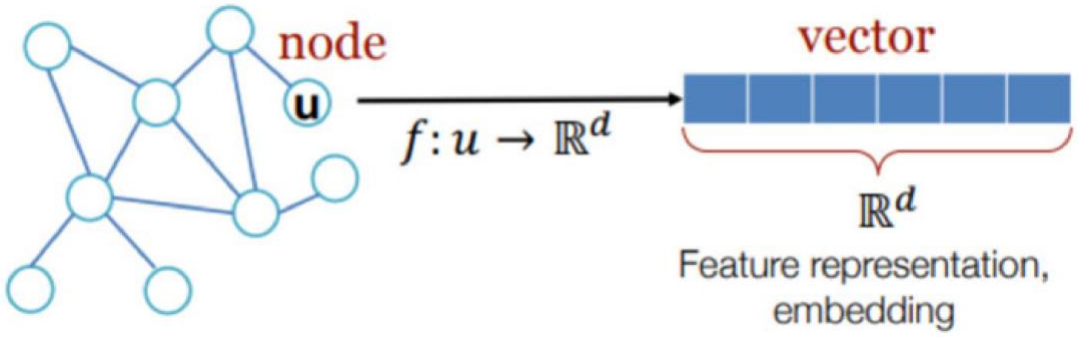


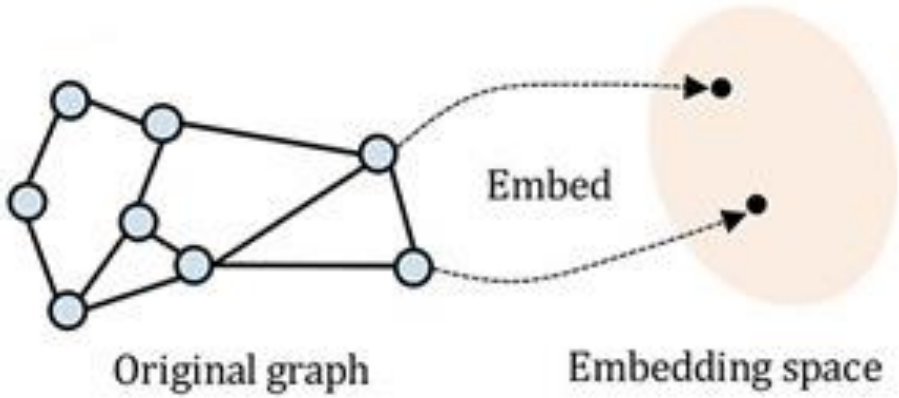
本节目录

- **图表示学习技术**
  - 基于随机游走的图表示学习
  - 基于图神经网络的图表示学习
- 知识图谱表示学习
- 知识图谱推理补全
- 补充：事件抽取概述

- **图表示学习问题概述**

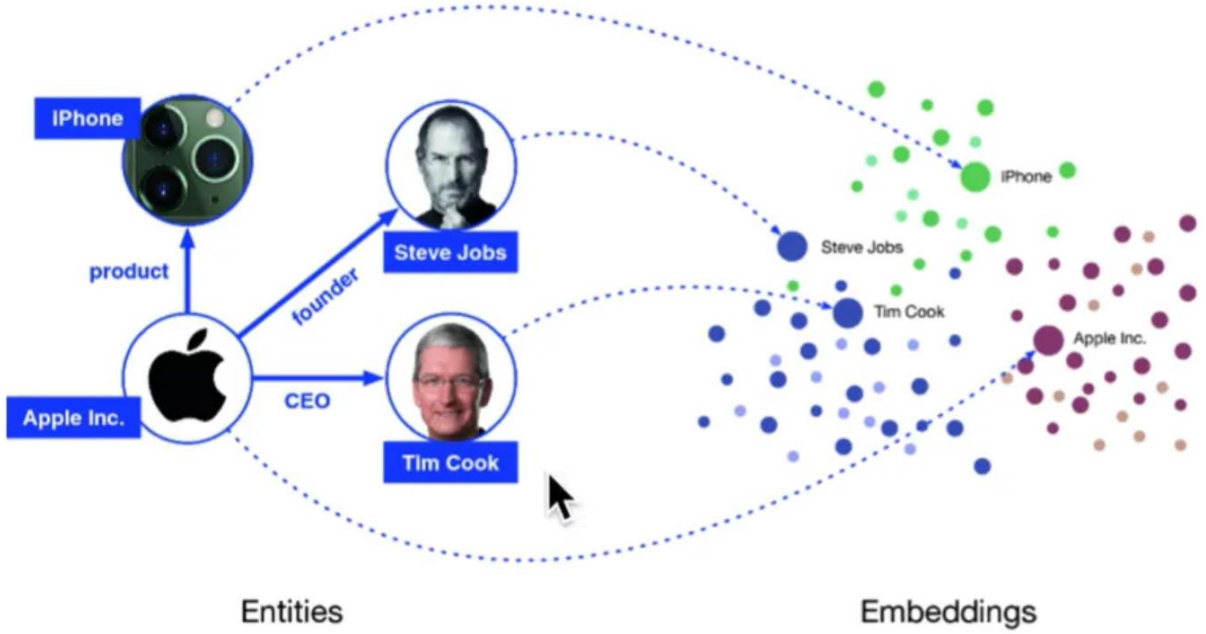
- 图表示学习算法
  - 将图数据进行向量化表征，映射到一个低维的向量空间，在这个低维向量空间中，图的结构特征和语义特征得到最大限度的保留。
    - 输入：图结构信息 $G =< V, E >$
    - 输出：向量化表征 $f: v_i \rightarrow Z_i \in \mathbb{R}^d, \quad d \ll |V|$





- **图表示学习问题概述**

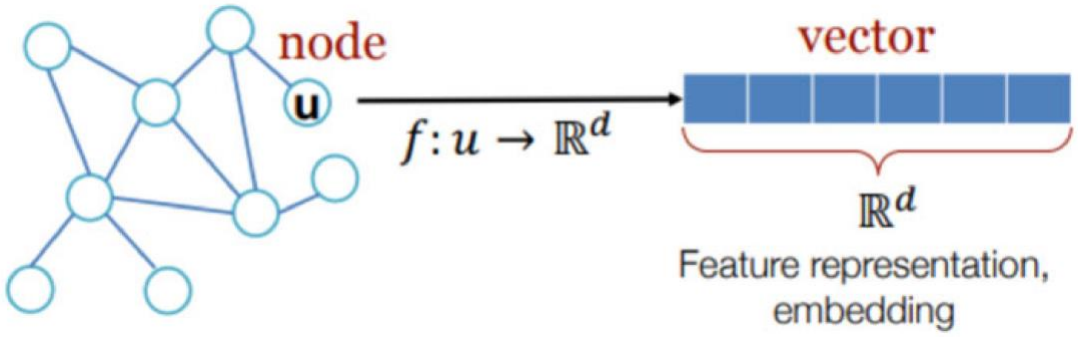
- 图表示学习算法的核心假设
  - 图表示学习需要找到某种节点相似度目标（Loss）以学习节点向量
    - 例如，语义特征相似节点的向量化表示应该是相似的

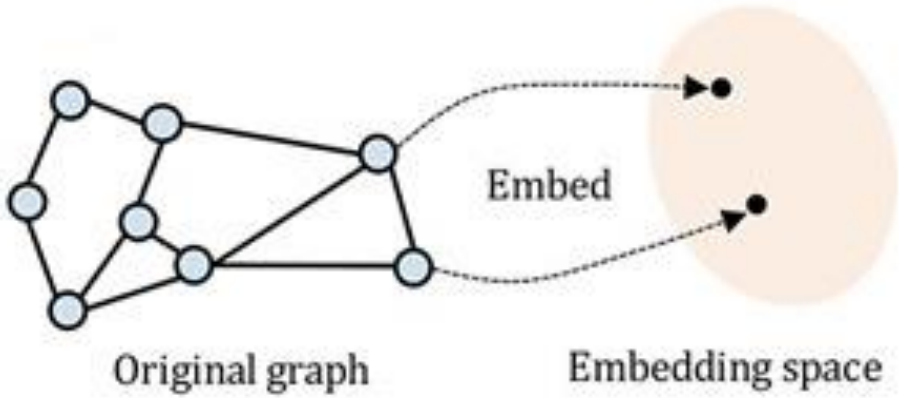


- **最基本的图表示学习**

➢ 图的基本表示形式之一：邻接图矩阵

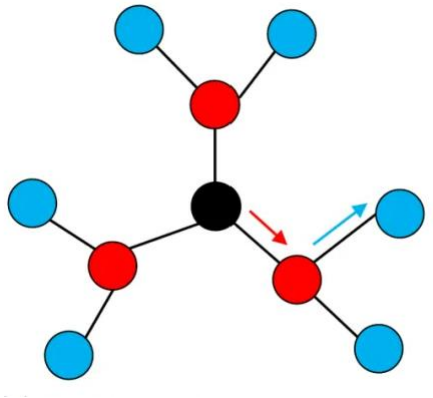
- 其中，每一行表示一个节点，1/0分别表示与对应节点是/否连接
- 启发式想法：这一行可以视作该节点的一个表示向量
  - 下游应用：该思路可用于最基础的图聚类问题
  - 局限性：未能充分融入节点结构信息，节点属性信息无法加入

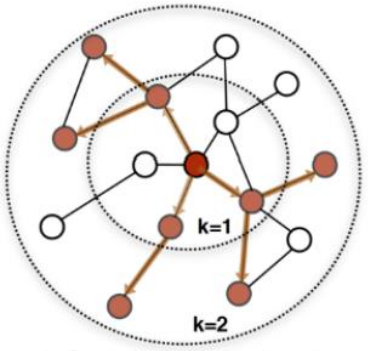


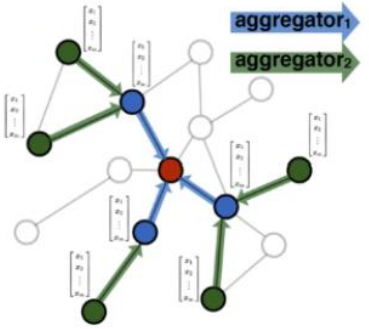


- **进阶：两种不同的图表示学习算法**

- 基于随机游走的图表示学习
  - 基于随机游走的邻居节点序列，挖掘图结构信息
  - ➢ DeepWalk、Node2vec、Metapath2vec…
- 基于图神经网络的图表示学习
  - 利用神经网络来学习图结构数据，提取和挖掘图结构中的特征和模式
  - ➢ GCN、GraphSAGE、GAT…







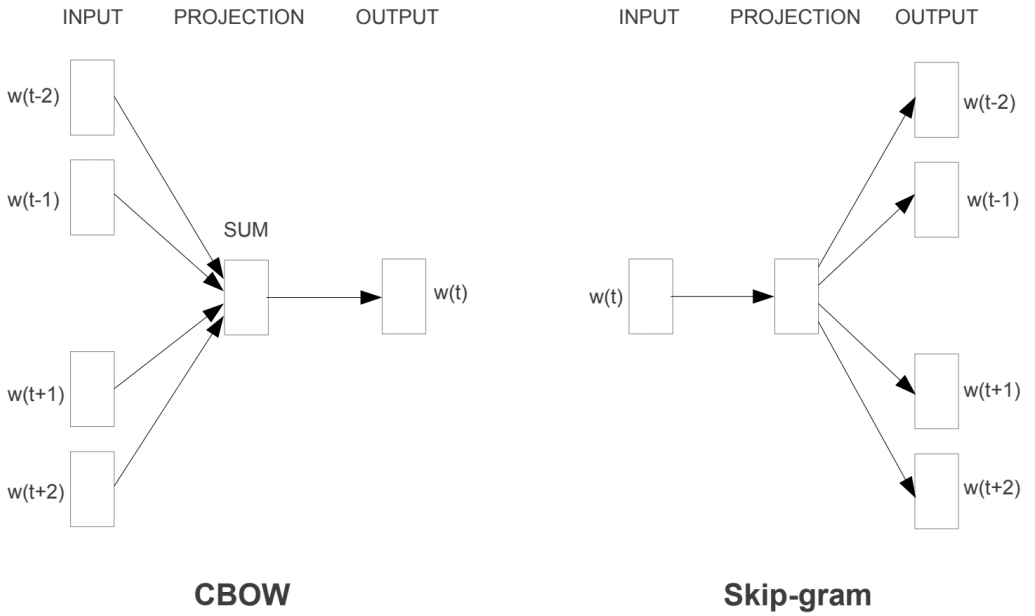
本节目录

- **图表示学习技术**
  - **基于随机游走的图表示学习**
  - 基于图神经网络的图表示学习
- 知识图谱表示学习
- 知识图谱推理补全
- 补充：事件抽取概述

- **回顾一下：NLP中如何处理序列**

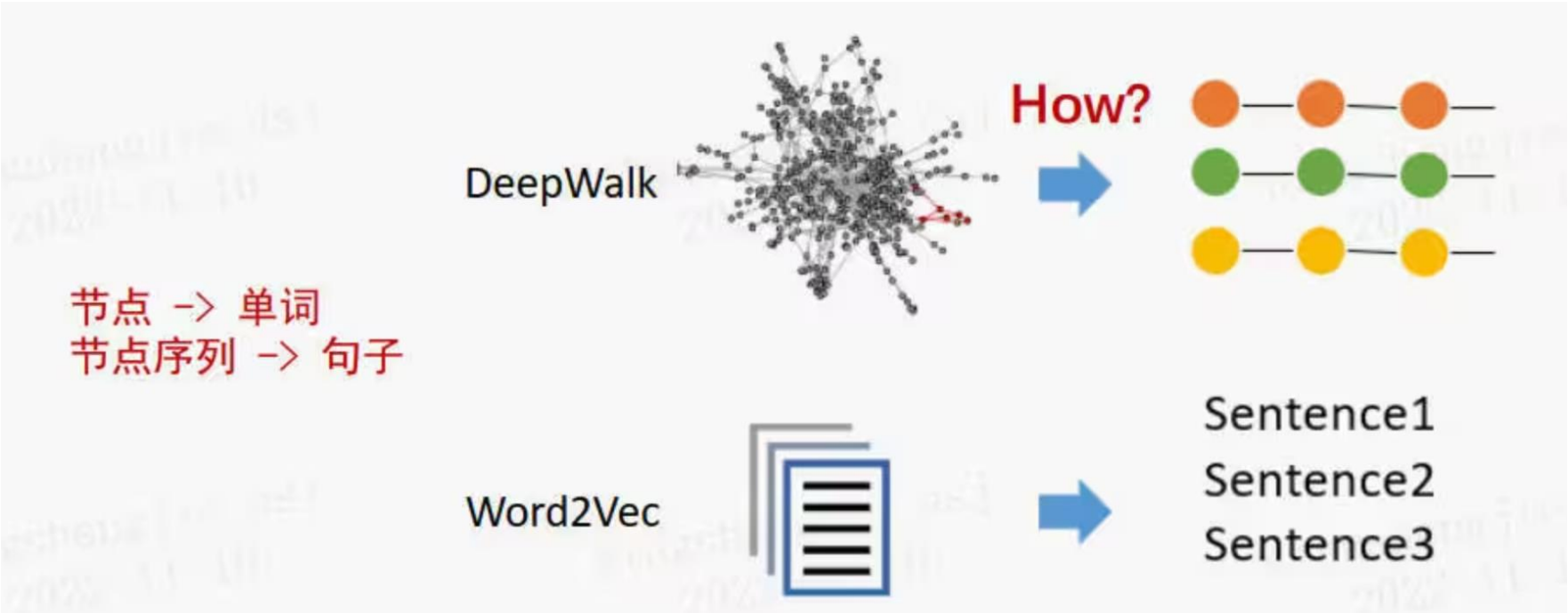
- 如何通过设计模型以及优化目标，来获得词项间更为一般化的关系表征
- Word2vec模型的出现：两种设计思路
  - 根据上下文预测中心词（CBOW）

  - 根据中心词预测上下文（Skip-gram）



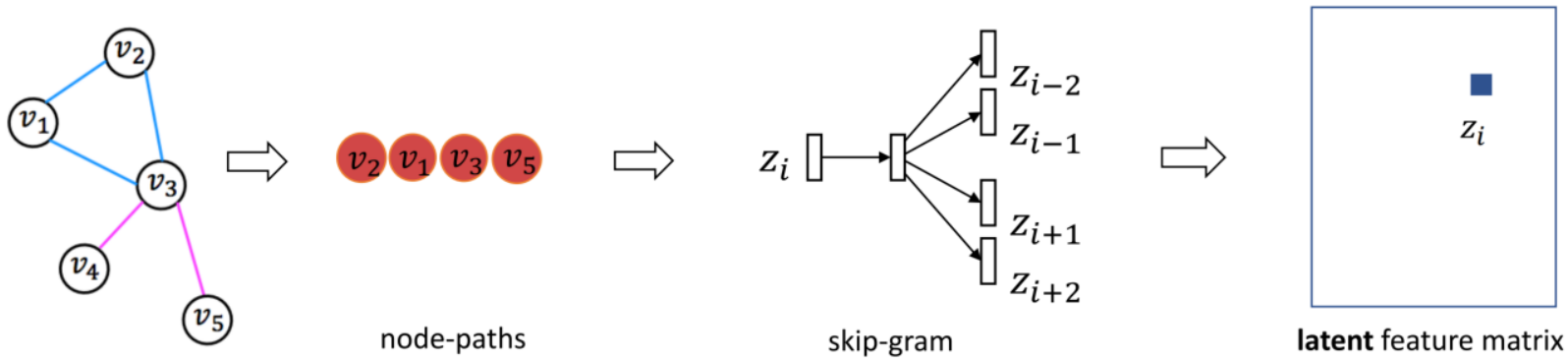
- **DeepWalk**

- 利用随机游走思想，通过随机选择相邻节点，获取节点序列
  - 游走的序列$Z = \{z_1, z_2, \dots, z_i, \dots z_l\}$可以类比语料库中的句子



- **DeepWalk**

- 采用skip-gram算法的思想，利用Word2Vec方式学习节点表征
  - 针对某个节点z_i，通过预测周围节点出现的概率学习节点z_i表征



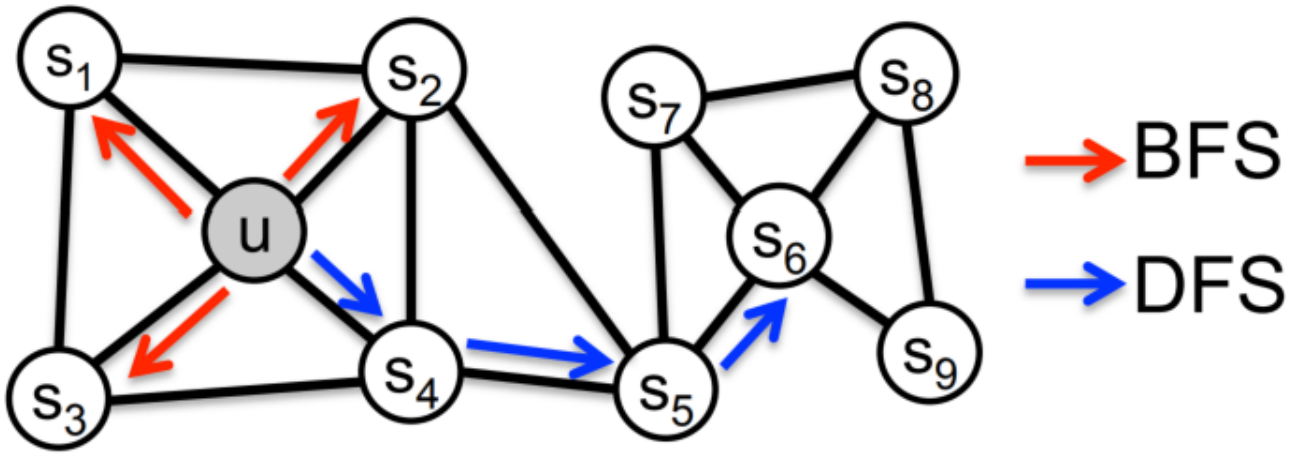
- 目标函数定义如下，采用无监督的方式训练模型

$$\Pr\left(\{v_{i-w},\cdots,v_{i+w}\}\setminus v_i \mid \Phi(v_i)\right) = \prod_{\substack{j=i-w \\ j\neq i}}^{i+w} \Pr(v_j|\Phi(v_i))$$

有关DeepWalk及相关图表示学习，可参考B站教程： https://www.bilibili.com/video/BV1o94y197vf

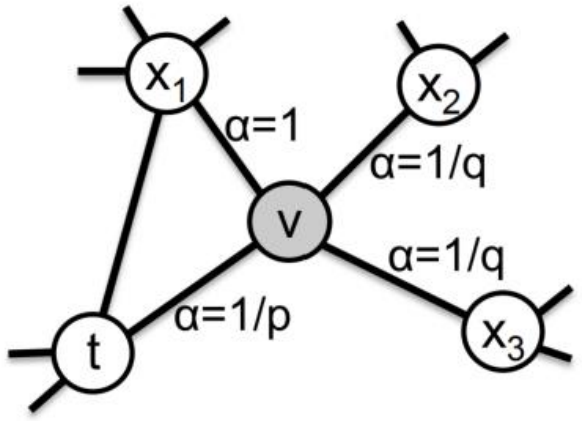
- **Node2vec**

- 可以看作DeepWalk算法的改进，使用有策略的游走算法产生节点序列。
  - 核心假设：homophily and structural equivalence
    - Homophily，表示宏观同质性，即相邻节点属性/表征相似，对应**DFS**
    - Structural Equivalence，表示局部结构相似（有共同邻居），对应**BFS**



- **Node2vec**

- 可以看作DeepWalk算法的改进，使用有策略的游走算法产生节点序列。
  - 游走策略
    - p决定多大概率往回走，即重复访问刚路过的节点（BPS）
    - q决定多大概率往远处走（DPS）



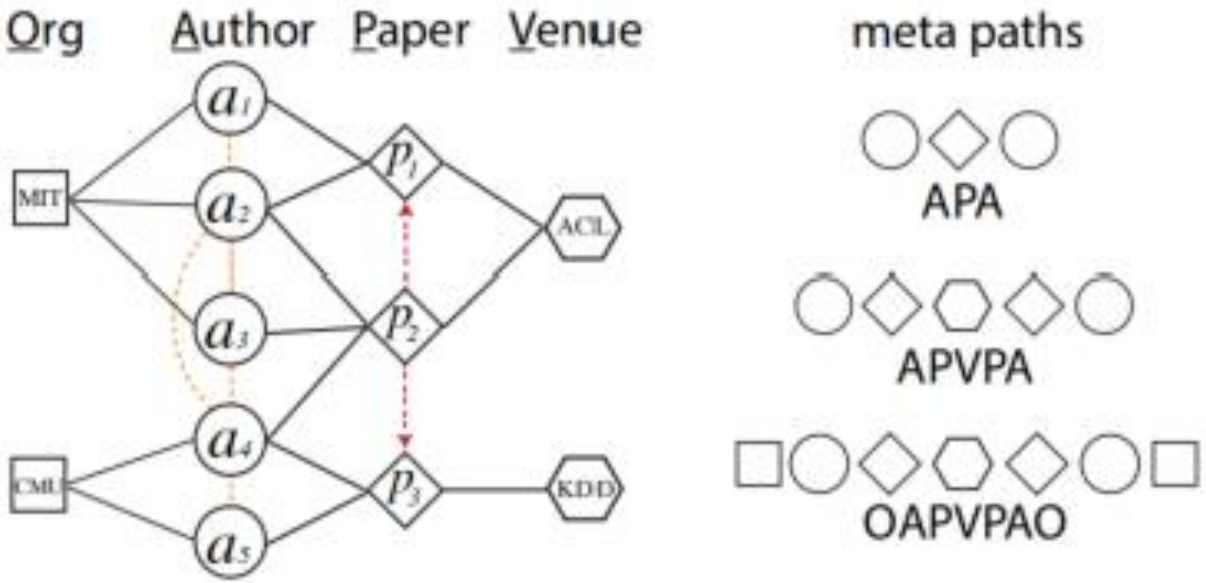
$$\alpha_{pq}(t,x) = \begin{cases} \frac{1}{p} & \text{if } d_{tx} = 0 \\ 1 & \text{if } d_{tx} = 1 \\ \frac{1}{q} & \text{if } d_{tx} = 2 \end{cases}$$

Dtx表示节点t与节点x之间的**最短距离**，因此

- Dtx=0表示回到t节点，即走回头路
- Dtx=2表示远离t节点

- **Metapath2Vec**

- 采用和前两种算法相同的核心思路：随机游走+skip-gram
- 但是，通过引入Meta-path修改随机游走方式可实现异质图的表示学习
  - 基于Meta-path游走产生节点序列（Meta-path事先人工定义）



- **Metapath2Vec**

- 通过引入Meta-path修改随机游走方式可实现异质图的表示学习
  - 基于Meta-path的随机游走策略

$$p(v^{i+1}|v_t^i,\mathcal{P})=\begin{cases}\frac{1}{|N_{t+1}(v_t^i)|} & (v^{i+1},v_t^i)\in E,\phi(v^{i+1})=t+1\\ 0 & (v^{i+1},v_t^i)\in E,\phi(v^{i+1})\neq t+1\\ 0 & (v^{i+1},v_t^i)\notin E\end{cases}$$

两个节点之间没有边，或有边但与MetaPath的限定不符，都不能游走下去

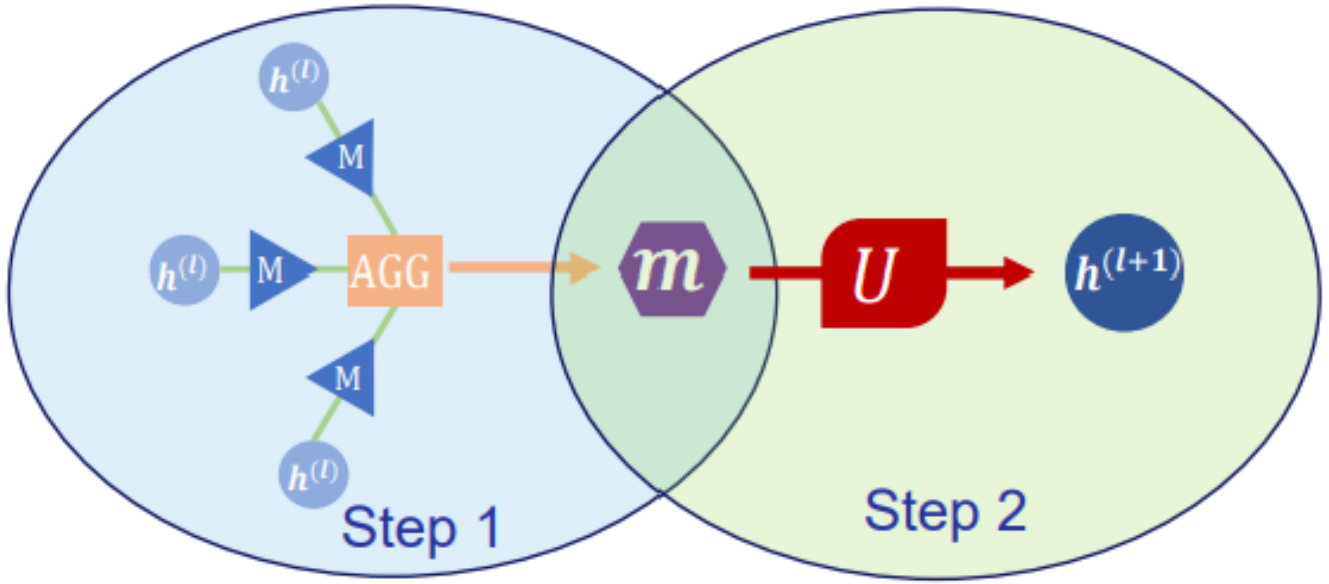
  - Metapath2Vec在类似2/K部图的场景上效果更佳
    - 例如，用户-商品，作者-论文-会议等

本节目录

- **图表示学习技术**
  - 基于随机游走的图表示学习
  - **基于图神经网络的图表示学习**
- 知识图谱表示学习
- 知识图谱推理补全
- 补充：事件抽取概述

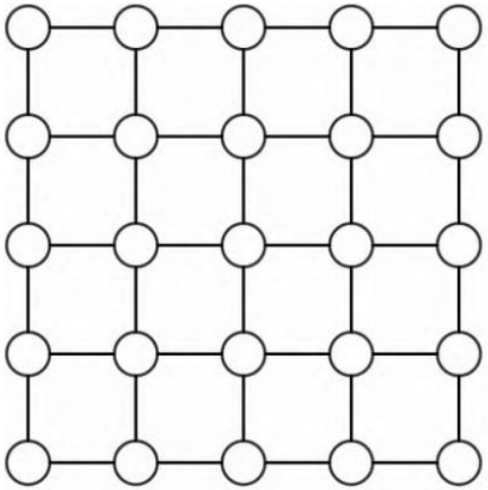
- **图神经网络技术**

- 从基本的路径游走，到面向表征整合的消息传递框架
  - 整合邻居节点的信息，在相邻节点间进行信息传播
  - 基于当前节点的表示和邻居节点的信息更新节点的表示

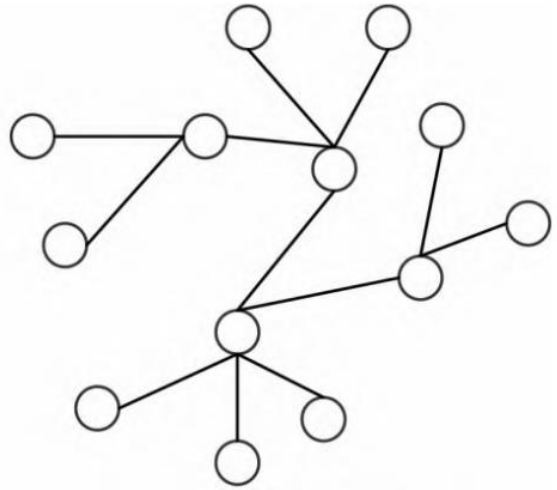


- **图卷积神经网络（Graph Convolutional Network, GCN）**

- GCN可视作图神经网络的开山之作
  - 传统神经网络技术（包括CNN）仅能处理常规欧式数据
  - 而图（Graph）属于典型非欧式数据，如何借助神经网络进行处理？



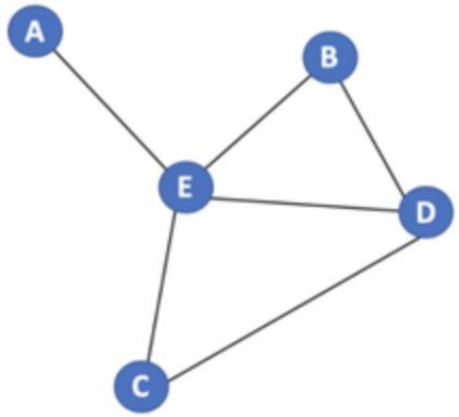
欧式数据



非欧式数据

- **图卷积神经网络（Graph Convolutional Network, GCN）**

- GCN可视作图神经网络的开山之作
  - 直观想法：可以将图转化为邻接矩阵，并借助矩阵形式进行处理



Graph *G*

	A	B	C	D	E
A	0	0	0	0	1
B	0	0	0	1	1
C	0	0	0	1	1
D	0	1	1	0	1
E	1	1	1	1	0

Adjacency matrix *A*

	A	B	C	D	E
A	1	0	0	0	0
B	0	2	0	0	0
C	0	0	2	0	0
D	0	0	0	3	0
E	0	0	0	0	4

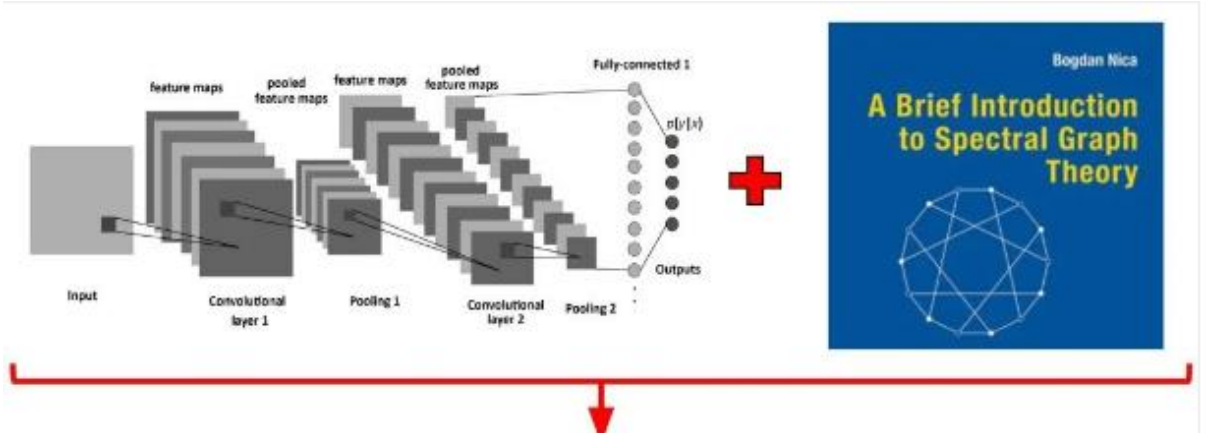
Degree matrix *D*

A	-1.1	3.2	4.2
B	0.4	5.1	-1.2
C	1.2	1.3	2.1
D	1.4	-1.2	2.5
E	1.4	2.5	4.5

Feature vector *X*

- **图卷积神经网络（Graph Convolutional Network, GCN）**

- 基于卷积神经网络知识和谱图理论推导出了节点的聚合函数形式



$$H^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)})$$

- 其中，A为邻接矩阵，D为度矩阵，W为神经网络的权重矩阵
- H为激活矩阵，其中H0 = X（样本特征向量）

完整的GCN演化史可以参考：https://zhuanlan.zhihu.com/p/120311352

• 图卷积神经网络（Graph Convolutional Network, GCN）

- 优点
  - GCN善于编码图的结构信息，能够学习到更好的节点表示。
  - 即使不训练，完全使用随机化参数，GCN就能够提取较好的图特征
- 缺点
  - GCN需要对整张图进行计算，很难应用在超大图上
  - GCN无法泛化新加入的节点，为新节点产生embedding需要额外的操作

- **Graph Sample and Aggregate（GraphSAGE）**

- 可解决前面所述GCN的两个缺点问题
  - 将整张图的学习优化到当前邻居节点的采样
  - Aggerator和权重矩阵的参数对于所有节点是共享的
  - 训练时仅保留训练样本到训练样本的边
- 有趣的是，GraphSAGE既可以有监督学习，也可以无监督学习（聚类）

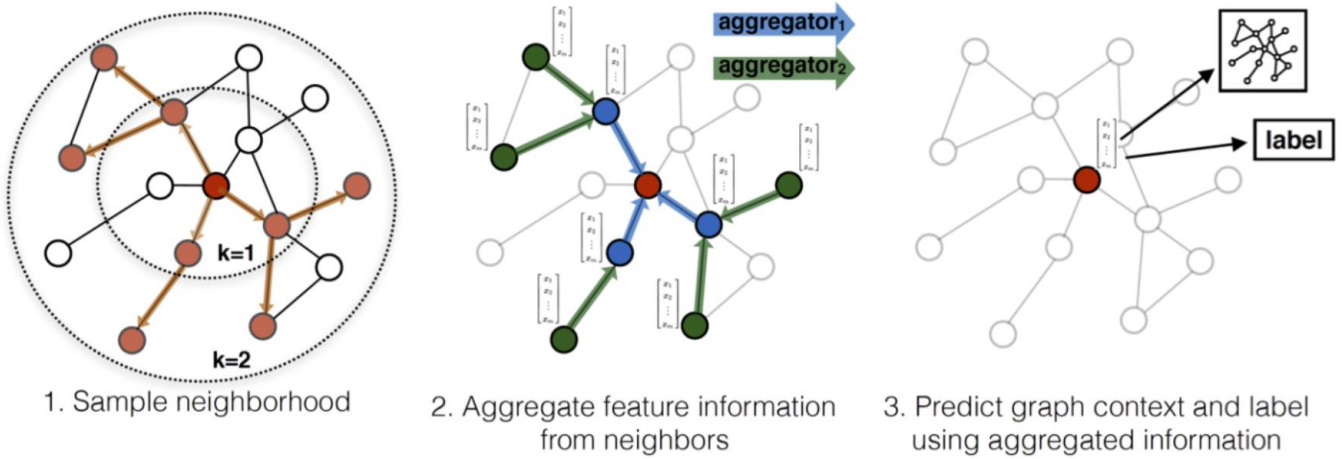


Figure 1: Visual illustration of the GraphSAGE sample and aggregate approach.

• Graph Sample and Aggregate（GraphSAGE）

Algorithm 1: GraphSAGE embedding generation (i.e., forward propagation) algorithm

Input : Graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$; input features $\{\mathbf{x}_v, \forall v \in \mathcal{V}\}$; depth K; weight matrices $\mathbf{W}^k, \forall k \in \{1, ..., K\}$; non-linearity σ; differentiable aggregator functions $\text{AGGREGATE}_k, \forall k \in \{1, ..., K\}$; neighborhood function $\mathcal{N} : v \rightarrow 2^{\mathcal{V}}$

Output : Vector representations $\mathbf{z}_v$ for all $v \in \mathcal{V}$

$\mathbf{h}_v^0 \leftarrow \mathbf{x}_v, \forall v \in \mathcal{V}$;
for $k = 1...K$ **do**
  for $v \in \mathcal{V}$ **do**
    $\mathbf{h}_{\mathcal{N}(v)}^k \leftarrow \text{AGGREGATE}_k(\{\mathbf{h}_u^{k-1}, \forall u \in \mathcal{N}(v)\})$;
    $\mathbf{h}_v^k \leftarrow \sigma\left(\mathbf{W}^k \cdot \text{CONCAT}(\mathbf{h}_v^{k-1}, \mathbf{h}_{\mathcal{N}(v)}^k)\right)$
  end
  $\mathbf{h}_v^k \leftarrow \mathbf{h}_v^k / \|\mathbf{h}_v^k\|_2, \forall v \in \mathcal{V}$
end
$\mathbf{z}_v \leftarrow \mathbf{h}_v^K, \forall v \in \mathcal{V}$

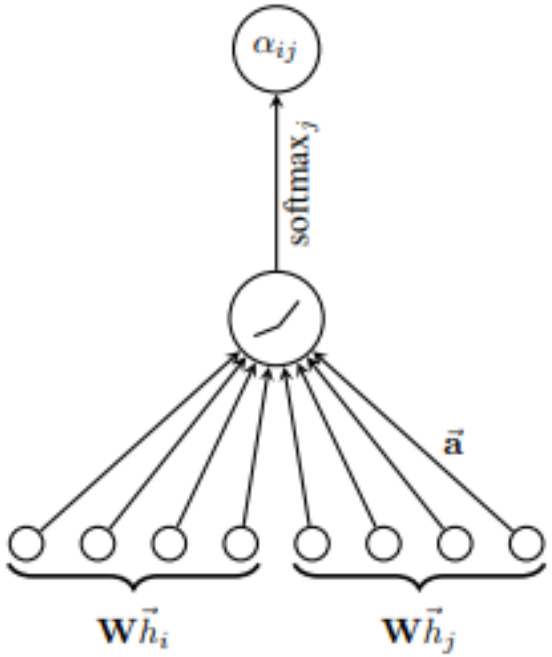
- Mean $\quad \text{AGG} = \sum_{u \in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|}$
- Pool $\quad \text{AGG} = \gamma\left(\{\mathbf{Q}\mathbf{h}_u^{k-1}, \forall u \in N(v)\}\right)$ （element-wise mean/max）
- LSTM $\quad \text{AGG} = \text{LSTM}\left([\mathbf{h}_u^{k-1}, \forall u \in \pi(N(v))]\right)$

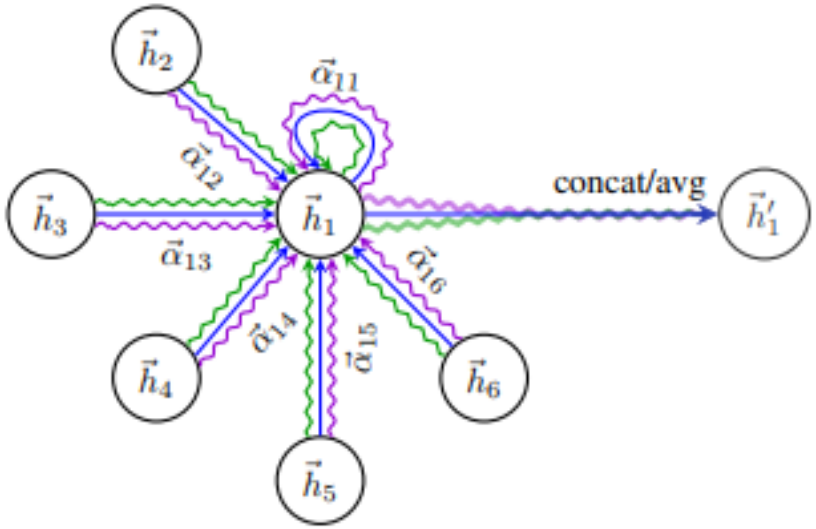
• Graph Sample and Aggregate（GraphSAGE）

- 优点
  - ➢ 采用邻居节点采样机制，克服了GCN训练时内存和显存的限制
  - ➢ Aggerator和权重矩阵的参数对于所有节点是共享的
- 缺点
  - ➢ 采样数目限制可能会导致部分节点的重要信息丢失
  - ➢ 采样引入随机化过程，没有区分不同邻居节点的重要性不同

- **图注意力神经网络（Graph Attention Network, GAT）**

- 可解决GraphSAGE没有考虑不同邻居节点重要性不同的问题
  - ➢ 带权重的聚合操作：为每个邻居节点分配不同的权值





$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_k]\right)\right)}$$

$$\vec{h}'_i = \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}\mathbf{W}\vec{h}_j\right)$$

GAT中注意力的计算方式可参考https://blog.csdn.net/qq_35091353/article/details/123896977

- **图注意力神经网络（Graph Attention Network, GAT）**

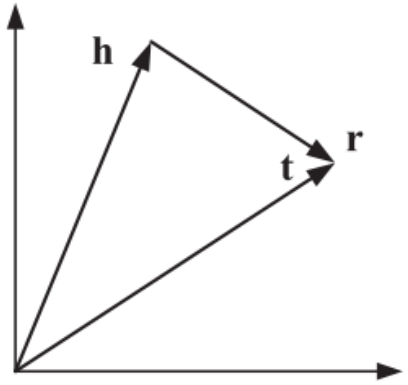
- 优点
  - 带权重的邻居节点聚合操作，可缓解随机邻居采样导致的同一节点embedding特征不稳定
  - 模型权值共享
  - 计算速度快，可以在不同的节点上并行计算
- 缺点
  - 由于邻居节点的高度重叠，导致冗余计算
  - 过平滑严重，导致隐层表征会收敛到同一个值，影响模型学习效果

本节目录

- 图表示学习技术
  - 基于随机游走的图表示学习
  - 基于图神经网络的图表示学习
- **知识图谱表示学习**
- 知识图谱推理补全
- 补充：事件抽取概述

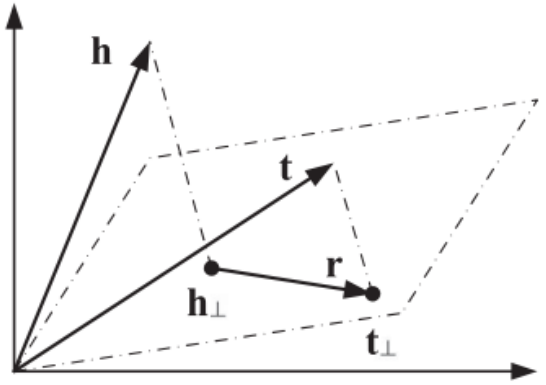
- **知识图谱+图学习：知识图谱表示学习**

- 为知识图谱中的实体和关系提供向量表示，同时保留语义信息
  - 不仅表征实体（节点），也表征关系（边）
  - 不仅表征结构信息，也融合富语义关系类别信息



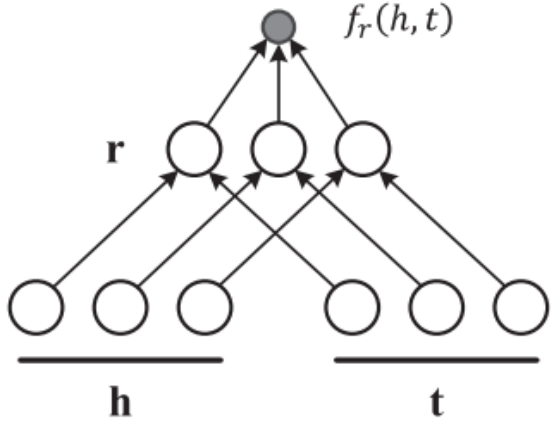
Entity and Relation Space

(a) TransE.



Entity and Relation Space

(b) TransH.



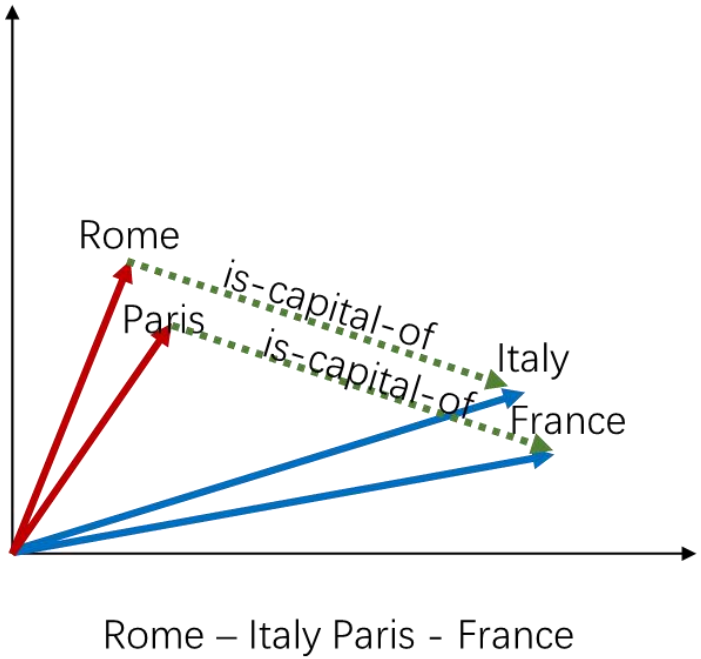
(b) DistMult.

- 下游任务：知识图谱推理补全　　应用场景：推荐系统、问答系统

• 知识图谱表示学习模型

• TransE

- 受 Word2vec 翻译不变性的启发： v(king)– v(queen) = v(man) – v(woman)



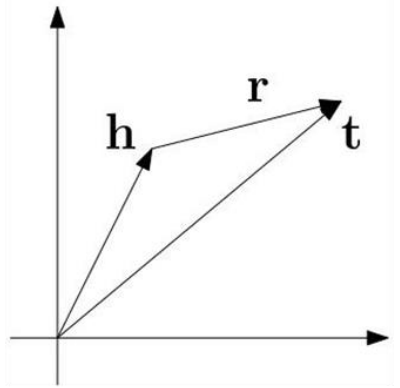
Rome – Italy Paris - France

$$\boldsymbol{h} + \boldsymbol{r} = \boldsymbol{t}$$

Rome + is-capital-of = Italy
Paris + is-capital-of = France

(Rome, is-capital-of, Italy)
(Paris, is-capital-of, France)

知识图谱中的三元组



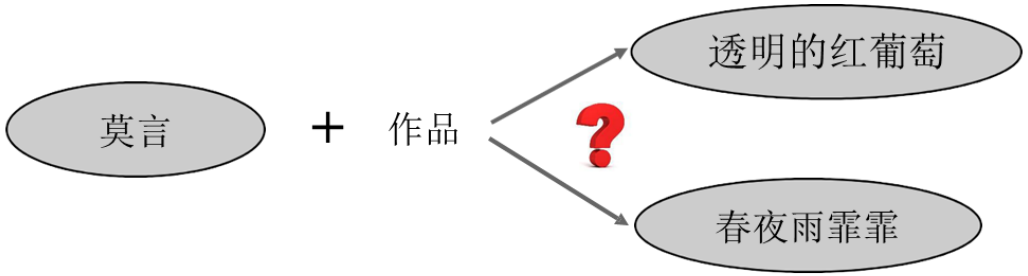
$$f_r(h,t) = \|h + r - t\|_{L_1/L_2}$$

$$J_{SE} = \sum_{t_r \in T_r} \sum_{t'_r \in T'_r} \max\left(0, [\gamma + f(t_r) - f(t'_r)]\right)$$

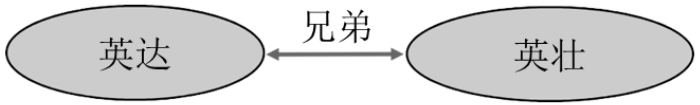
Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[J]. **NIPS**, 2013, 26.

- **知识图谱表示学习模型**

- **TransE 的问题**
  - 无法处理一对多、多对一和多对多问题



关系的性质



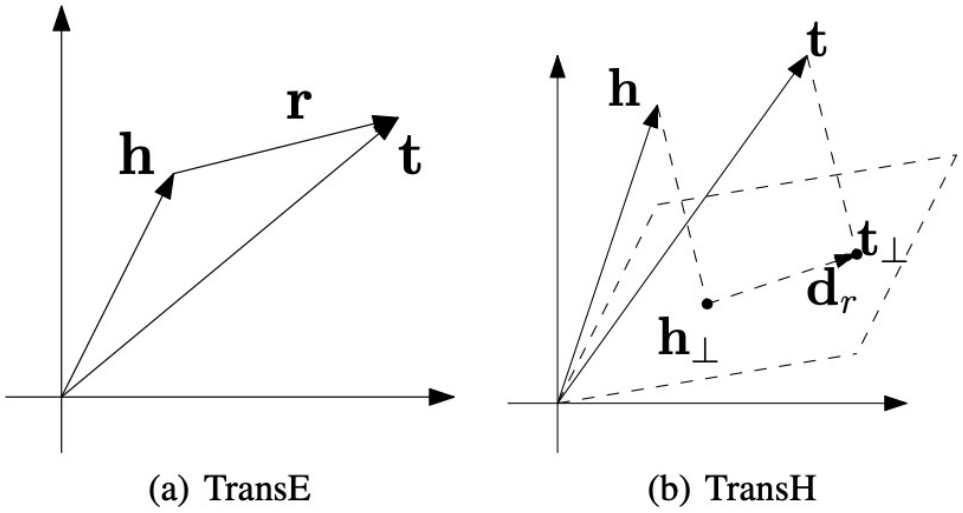
- **知识图谱表示学习模型**

- **TransH**

  - 通过向量 w_r 将实体投影到关系 r 对应的超平面上，解决一对多问题

$$\mathbf{h}_\perp = \mathbf{h} - \mathbf{w}_r^\top \mathbf{h}\mathbf{w}_r, \quad \mathbf{t}_\perp = \mathbf{t} - \mathbf{w}_r^\top \mathbf{t}\mathbf{w}_r.$$

$$f_r(h,t) = \|\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp\|_2^2,$$



(a) TransE (b) TransH

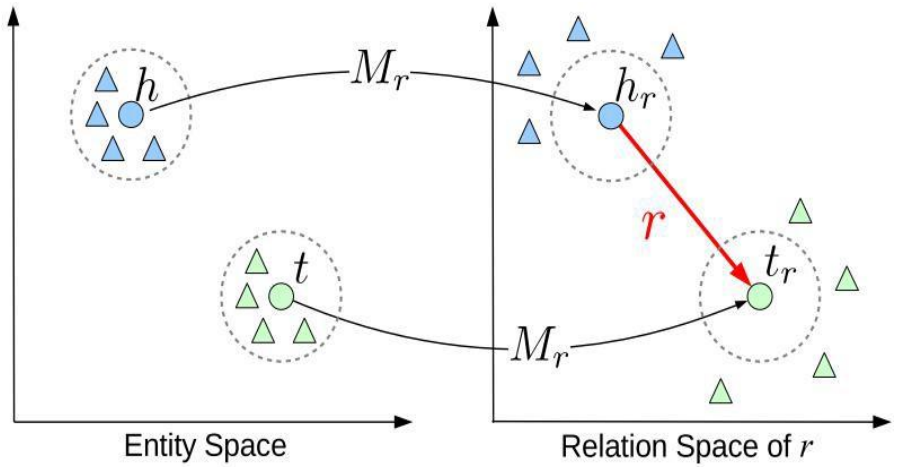
Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes **AAAI**. 2014

- **知识图谱表示学习模型**

- **TransR**
  - 通过矩阵 M_r 将实体进行坐标转换到关系 r 对应的关系空间中
    - 不同的关系对应实体不同的属性

$$\mathbf{h}_r = \mathbf{h}\mathbf{M}_r, \quad \mathbf{t}_r = \mathbf{t}\mathbf{M}_r.$$

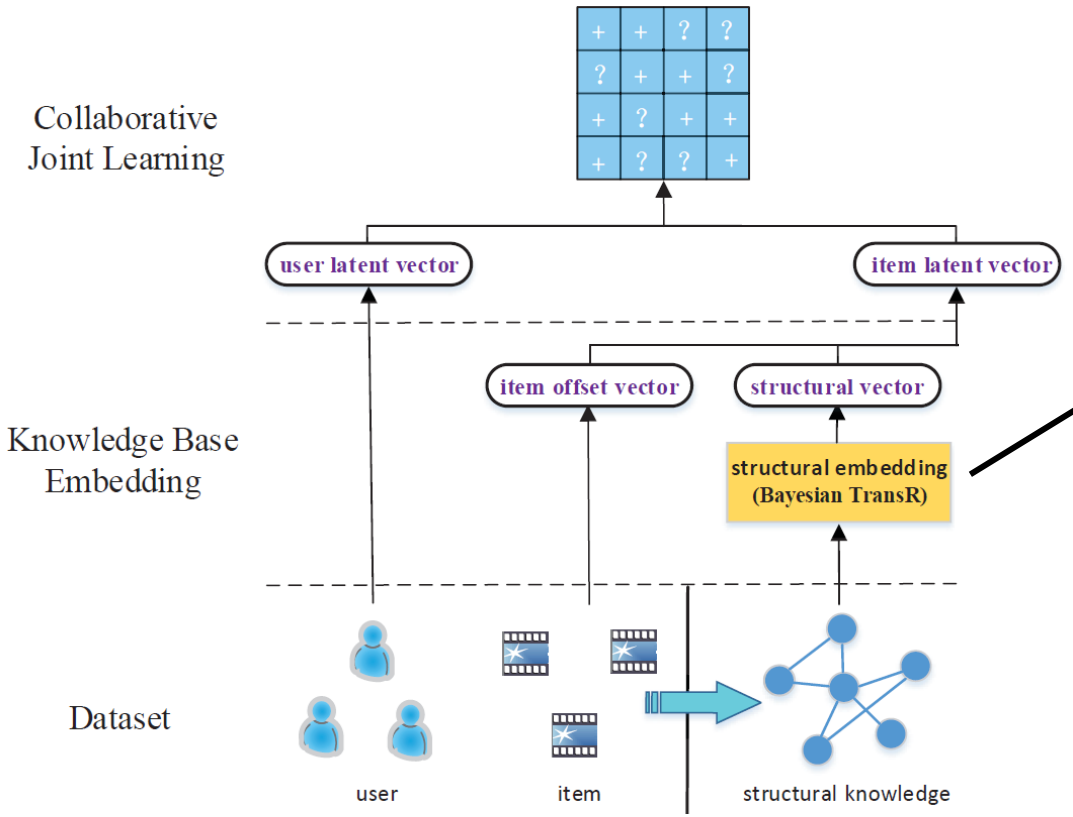
$$f_r(h,t) = \|\mathbf{h}_r + \mathbf{r} - \mathbf{t}_r\|_2^2.$$

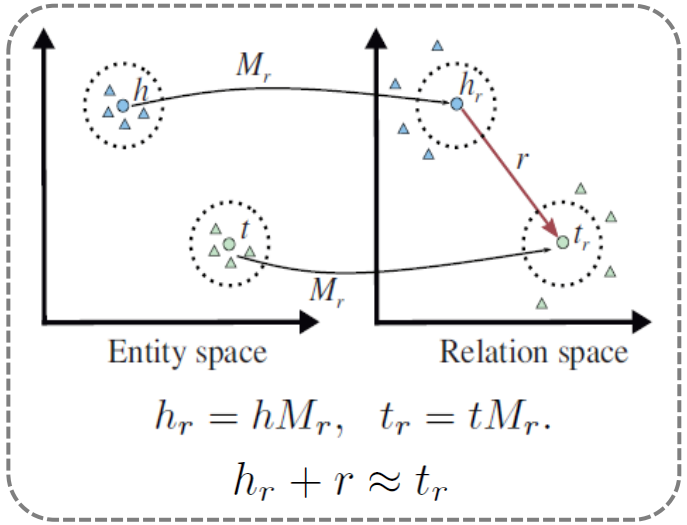


Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion **AAAI**. 2015.

- **拓展：基于图谱推理的推荐技术**

- 代表性方法：CKE (Zhang F, et al. "Collaborative knowledge base embedding for recommender systems." KDD 2016)





$\mathbf{e}_i = \boldsymbol{\eta}_i + \mathbf{v}_i$ ← Vi即基于图谱的Item表征

$\hat{y}_{ui} = \boldsymbol{\eta}_u^\top \mathbf{e}_i$

- **知识图谱表示学习模型**

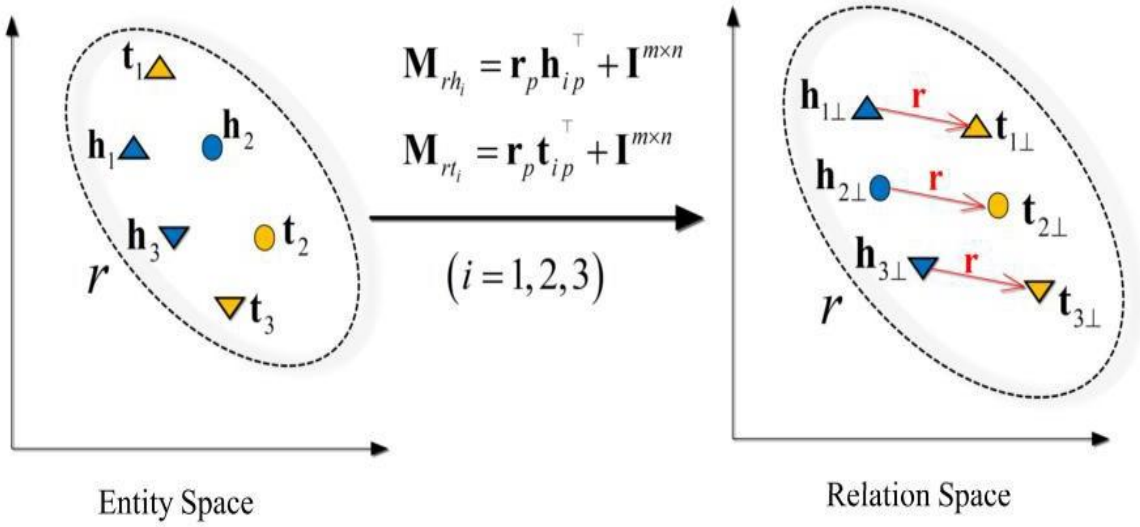
- **TransD**

  - 关系 r 对应的投影矩阵是动态生成的，描述同一种关系的不同语义

$$f_r(h,t) = \|M_{rh}h + r - M_{rt}t\|$$

$$M_{rh} = r_p h_p^T + I^{m\times n}$$

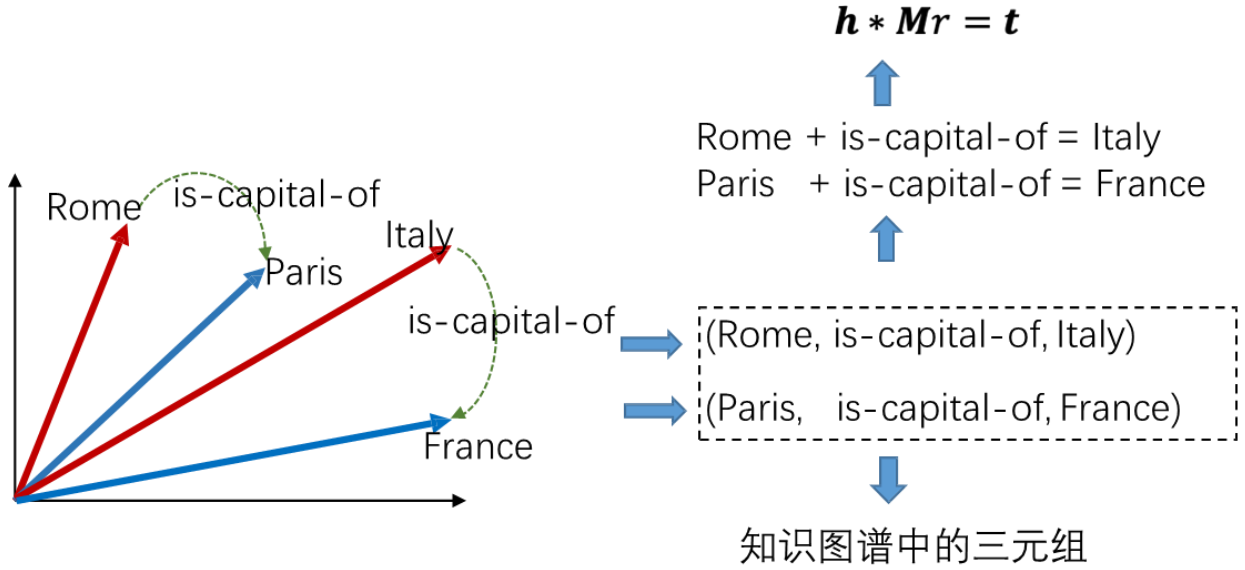
$$M_{rt} = r_p t_p^T + I^{m\times n}$$

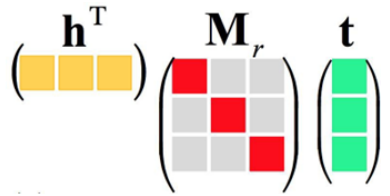


Ji, Guoliang, et al. Knowledge graph embedding via dynamic mapping matrix **ACL**. 2015.

- **知识图谱表示学习模型**

- **DistMult**
  - 每个关系 r 都被表示为一个矩阵 M_r，其建模了潜在因子间的成对相互作用
  - 将 M_r 限制为对角矩阵以简化计算（本质上是把关系当做线性变换）





线性变换假设：

$$hM_r = t$$

目标函数：

$$f_r(h,t) = hM_rt$$

Yang B, YihW, He X, et al. Embedding entities and relations for learning and inference in knowledge bases **ICLR**, 2014.

本节目录

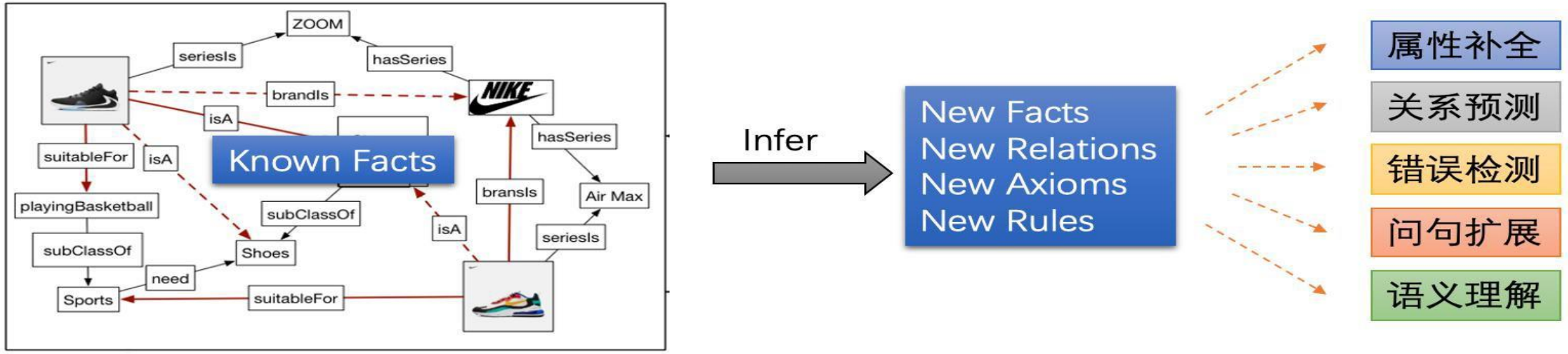
- 图表示学习技术
  - 基于随机游走的图表示学习
  - 基于图神经网络的图表示学习
- 知识图谱表示学习
- **知识图谱推理补全**
- 补充：事件抽取概述

- **什么是推理?**

- 从 已知事实 推断出 新的事实或知识 的过程
  - 演绎
    - 工作日要上班 + 今天是工作日 ⟶ 今天要上班
  - 溯因:
    - 下雨草地会湿 + 观察到草地湿了 ⟶ 可能下雨了
  - 归纳:
    - 以往观察到的天鹅都是白色的 ⟶ 所有的天鹅都是白色的
  - 类比:
    - 菠萝饼干与草莓饼干的做法应该差不多

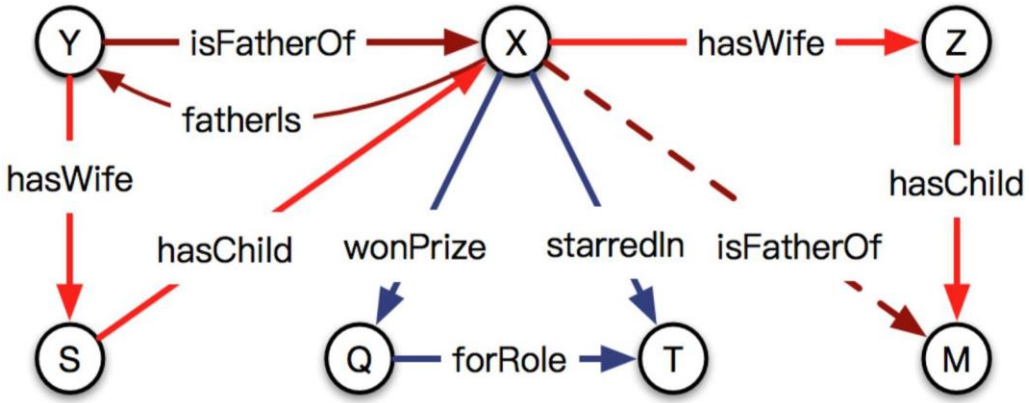
- **知识图谱推理**

- 主要关注围绕关系的推理
  - 基于图谱中已有的事实或关系来推断未知的事实或关系
- 许多现实中的问题（如链接预测、因果推理、基于知识图谱的问答和推荐等）都可以表述为图谱推理



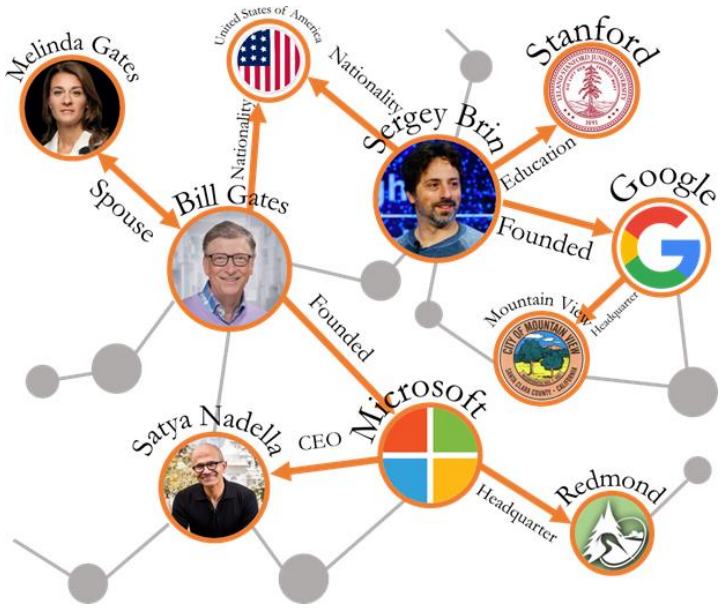
- **知识图谱上的关系推理任务**

- 链接预测
  - 给定两个实体，预测它们之间是否存在 r 关系
- 实体预测
  - 给定头实体和关系，预测未知的尾实体
- 事实三元组预测
  - 给定一个三元组判断其是否为真或假



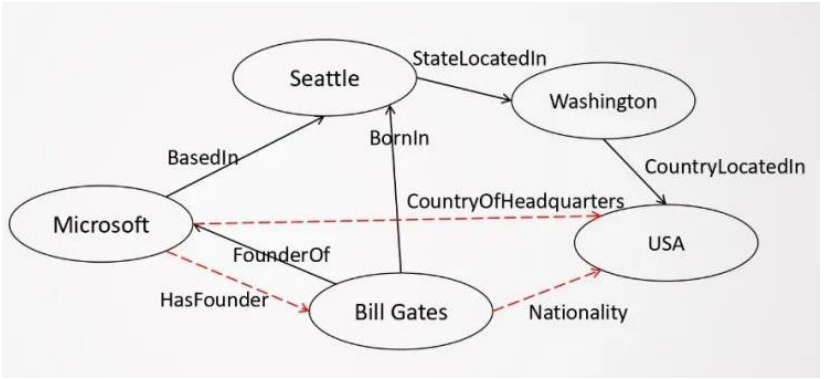
- **相似的任务：关系补全**

- 知识图谱“生来”不完整，多数的现有知识图谱都是稀疏的，由此引出了图谱补全来向知识图谱添加新的三元组
  - 我们需要做的，是基于图谱里已有的事实，去推理出缺失的事实
  - 本质上说，图谱补全和推理任务非常类似
    - 都是“无中生有”的过程
    - 推理的结果可以作为新“知识”加入图谱



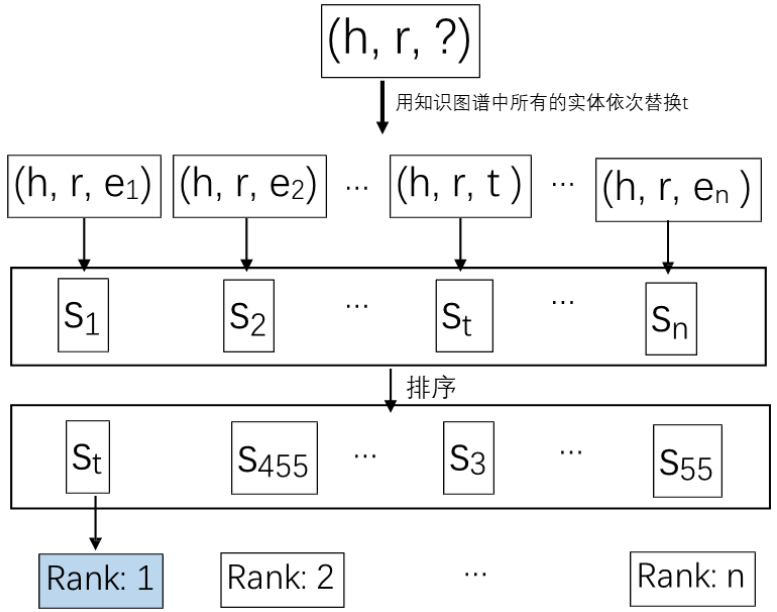
- **图谱补全的任务是什么?**

- 给定知识图谱G = {E, R, F}，其中E表示所有实体的集合，R表示所有关系的集合，F为所有三元组的集合
- 知识图谱补全的任务是预测出当前知识图谱中缺失的三元组F'={(h, r, t) | (h, r, t)∉F, r∈R}
  - 具体子任务包括头、尾实体预测和关系预测：
    - 给定三元组(_,r,t)，预测头实体
    - 给定三元组(h,r,_)，预测尾实体
    - 给定三元组(h,_,t)，来预测关系



- **知识图谱推理/补全的任务评价方式**

- 预测问题与推理评价
  - Hit@n：所有预测样本中排名在n以内的比例，n常用的取值为1，3，10
  - MR：Mean Rank 所有预测样本的平均排名
  - MRR：Mean Reciprocal Rank 先对所有预测样本的排名求倒数，然后求平均



- **知识图谱推理/补全方法论的分类**

- 基于符号逻辑的知识图谱推理
  - 显式的知识表示方法，一般需要人工定义
  - 可解释性好，但能进行的推理相对有限
- 知识图谱表示学习
  - 易于捕获隐式知识
  - 对机器友好，但不利于人理解

- **基于符号/规则逻辑的推理方法**

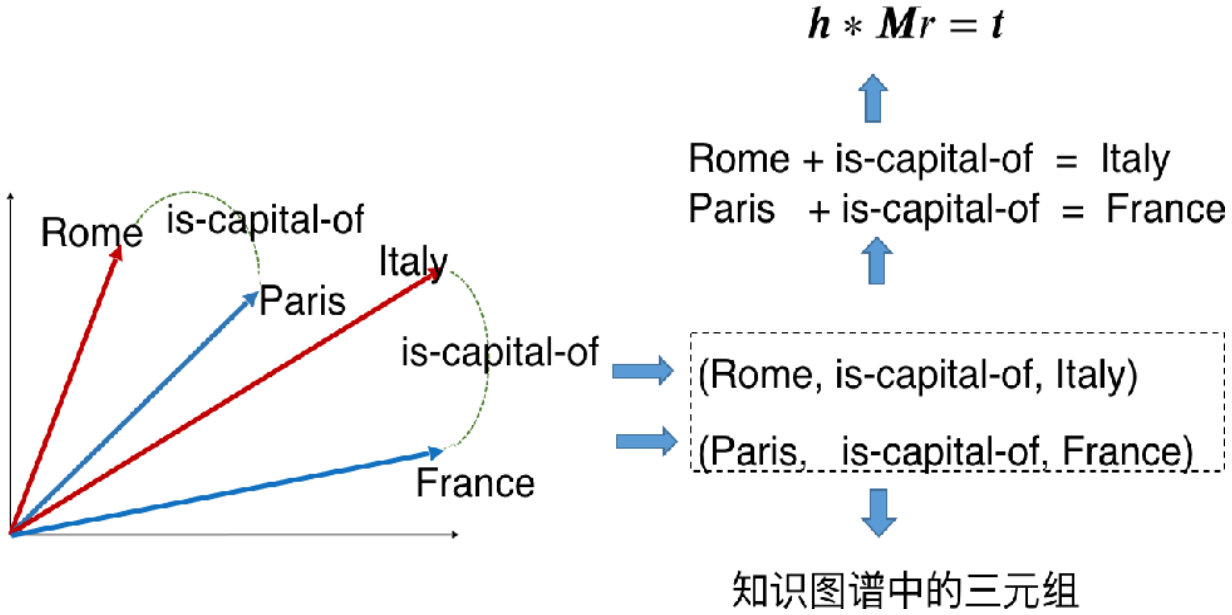
- KGC的一个研究方向是逻辑规则学习。规则由head和body以head←body的形式定义。头部是一个原子，而身体可以是一组原子。
  - 例如，给定关系sonOf，hasChild和sex以及实体X和Y，则有规则：

    （Y，sonOf，X）←（X，hasChild，Y）∧（Y，性别，男性）
  - 可以通过关联规则挖掘工具来提取逻辑规则
  - 但是目前更多的工作尝试将逻辑规则信息作为额外信息获得嵌入特征

- 有关符号推理部分的知识，感兴趣的同学可参考陈华钧老师的《知识图谱》课程

- **基于表示学习（嵌入）的归纳推理**

- 当有大量的关系或三元组需要推理时，基于向量表征的推理更有效率
  - 将图谱推理问题转化为向量计算问题
  - 缺陷：可解释性问题，即我们知道预测的结果，但不知道为什么



- **基于表示学习（嵌入）的归纳推理**

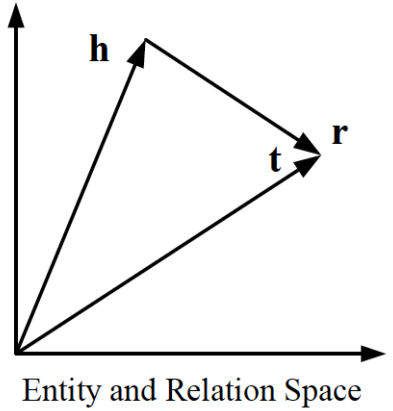
- 目前主流的研究方向是基于KG表征的补全方法
  - 基本框架为Encoding model +scoring function。
- Encoding Model部分：对于每个三元组(h,r,t)，通过encoding model学习实体和关系的特征表示
  - Encoding model可以是线性模型，也可以是神经网络模型等
  - 先前所介绍的各类知识图谱表示学习手段均可用于解决该任务

- **基于表示学习（嵌入）的归纳推理**

- 目前主流的研究方向是基于KG表征的补全方法
  - 基本框架为Encoding model +scoring function。
- Scoring function部分：结合标注数据，通过有监督学习训练模型，并评估新构造的三元组是否具有高的合理性
  - 真实的 fact 的score要比不真实的 fact 的score更高（回忆一下TransE）
  - 一般可以分为两类：基于距离、基于相似度

- **基于表示学习（嵌入）的归纳推理**

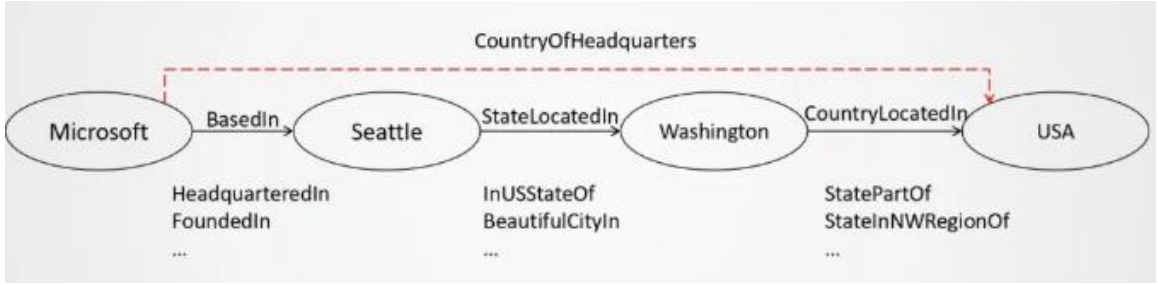
- 前面介绍的TransE使用的scoring function就是基于距离的，它认为头部的特征表示加上关系的特征表示应该就能得到尾部的特征表示
  - 而DisMult因为采用的是线性变换，使用的是基于相似度的方法
- 尽管在一些基准测试中，对实体和关系的表征学习已获得了显著的性能，但是它无法为复杂的关系路径建模



$$J_{SE} = \sum_{t_r \in T_r} \sum_{t'_r \in T'_r} \max\left(0, [\gamma + f(t_r) - f(t'_r)]\right)$$

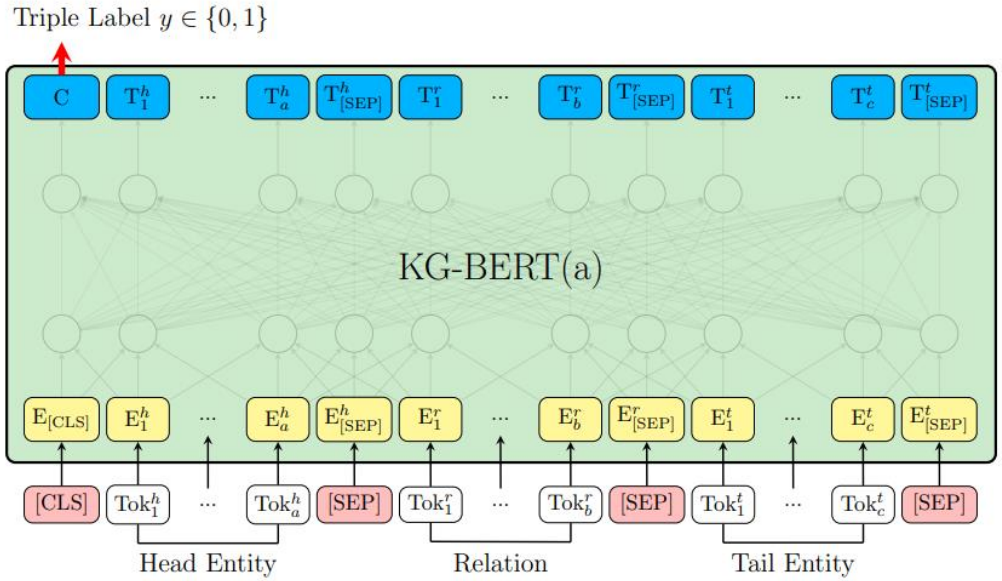
- **多跳图谱推理/补全方法**

- 基于路径查找的方法（**Pathfinding-based**）可以用来处理多步推理问题
  - 传统的路径查找方法主要是PRA方法（Path Ranking Algorithm）；但是这种方法对于包含较大规模的知识图谱来说，会由于路径数量爆炸式增长，导致特征空间急剧膨胀
  - 解决的方式，可以尝试用embedding的方式表示关系，对关系进行泛化，并基于此对知识的补全进行建模，以缓解路径数量过多导致的特征空间膨胀问题



- **引入大规模预训练模型**

- 把常规的图谱三元组补全问题纳入预训练语言模型的处理框架中来，将知识图谱补全的任务视作序列分类任务
  - 为了让预训练模型适应图谱中的fact，将模型在具体任务上继续fine-tuning
  - 对于**头、尾实体**补全问题：
    - 将三元组转换为文本序列模式
    - 获得[CLS]的表征向量并进行二分类，判断该三元组是否成立

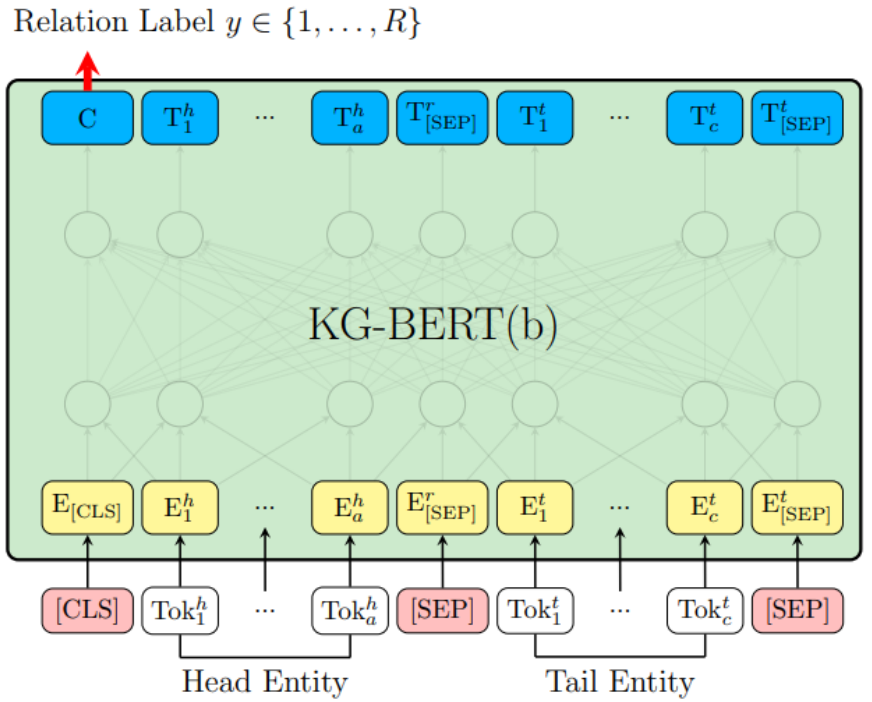


Yao L, Mao C S, KG-BERT: BERT for Knowledge Graph Completion

- **引入大规模预训练模型**

- 把常规的图谱三元组补全问题纳入预训练语言模型的处理框架中来，将知识图谱补全的任务视作序列分类任务
  - 对于**关系**补全问题：
    - 模型也可以完成关系分类，输入两个实体或实体描述，输出多类分类别
    - 采用负对数损失函数

$$\mathcal{L}' = -\sum_{\tau \in \mathbb{D}^+} \sum_{i=1}^{R} y'_{\tau i} \log(s'_{\tau i})$$



Yao L, Mao C S, KG-BERT: BERT for Knowledge Graph Completion

本节目录

- 图表示学习技术
  - 基于随机游走的图表示学习
  - 基于图神经网络的图表示学习
- 知识图谱表示学习
- 知识图谱推理补全
- **补充：事件抽取概述**

- **信息抽取的基本任务**

- 场景模板ST（事件抽取）
- 又称事件，是指实体发生的事件
  - 例如：会议 (Time<…>, Spot<…>, Convener<…>, Topic<…>)
  - 常见的新闻事件描述模板 5W1H
    - Who 、When 、Where 、What 、Why 、How
    - 例如：那天（When），有一个人（Who）来水果摊（Where）买瓜，因为秤有问题（Why），用西瓜刀（How）“*萨日朗*”（What）了郝哥

事件抽取的概念

- 事件是信息的一种表现形式，其定义为特定的人、物，在特定时间和特定地点相互作用所产生的客观事实。
  - 例如，可对应先前所说的5W1H基本要素
  - 一般信息呈现为句子级别（相比之下，关系往往表现为短语级别）
  - 推理任务（事件时序推理、事理推理）的前提
  - ACE中对于事件的定义如下：
    - *An event is a specific occurrence involving participants. An event is something that happens. An event can frequently be described as a change of state.*

- **事件抽取的基本要素**

- 通常而言，事件往往包含以下基本要素：
  - 事件触发词：表示事件发生的核心词，多为动词或名词
    - 相应的，事件触发词的检测与分类是事件抽取的基本任务

Example:

- *Henry[argument] was injured, and then passed away soon*

Detection: injured
Typing: Injure
Argument: Henry

Detection: passed away
Typing: Die
Argument : Henry

- **事件抽取的基本要素**

- 通常而言，事件往往包含以下基本要素：
  - 事件类型：与触发词相对应，往往可以通过触发词分类加以识别
    - 例如，前例中的触发词Pass away对应着“死亡”的事件类型
  - 事件元素：事件的参与者，主要由实体、时间等组成。
    - 例如，前例中的Henry是事件的主体
  - 事件元素角色：事件元素在事件中充当的角色。
    - 例如，前例中的Henry在事件中是一个“受害者”的角色

- **事件抽取的模板**

- 通过触发词识别和分类，判定事件及其类型后，可以借助模板实现抽取。
  - 我们曾提到过**模板元素TE**，其目的在于更加清楚、完整地描述实体
  - 其中，模板元素通过槽（Slots）描述了命名实体的基本信息
    - 槽的内容可包括名称、类别、种类等，不同类型的事件，对应的模板也不尽相同
    - 通过事件元素与元素角色的识别，将元素填入模板合适的槽，即完成了事件抽取

模版元素	实体类型	描述
Person-Arg	PER	结婚的人
Time-Arg	TIME-within	结婚时间
Place-Arg	GPE LOC FAC	结婚地点

- **事件抽取的模板**

- 在选定相应的模板之后，通过事件元素与事件元素角色的识别，将相应的元素填入模板合适的槽（Slot）内，即完成了事件抽取。
  - 案例：刚才有个朋友问我，马老师发生甚么事了
    - 元素/描述：人物（朋友），时间（刚才），事件（提问发生了甚么事）

- **限定域事件抽取**

- 某种意义上，限定域事件抽取与预定义关系抽取存在相似之处，即预先定义好目标事件的类型及每种类型的具体结构（包含哪些具体的事件元素）。
- 因此，限定域事件抽取可以采用基于模式匹配的方法实现
  - 可以采用完全规则的方法实现，即完全通过人工标注方式获得模式
  - 也可以采用弱监督的模式匹配，即不需要对语料进行完全标注，只需要人工对语料进行一定的预分类或者制定少量种子模式
    - 类似于前述的DIPRE方法，迭代式获得更完善的语料和模板

- **限定域事件抽取**

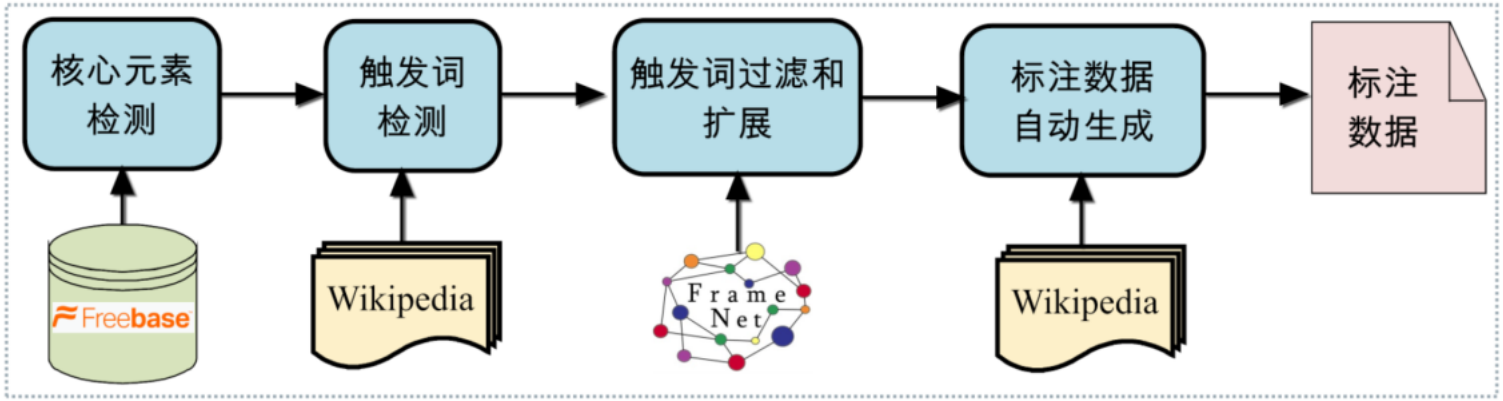
- 同样的，限定域事件抽取也可采用基于机器学习的方法实现。
- 例如，采用有监督学习方式，将事件抽取转化为一个多分类问题（传统艺能）
  - 基于特征工程的方法：将事件实例转换成分类器可以接受的特征向量
  - 基于神经网络的方法：自动从文本中获取特征进而完成事件抽取
  - 分类之后，不要忘记抽取要素填事件槽！
    - 当然，也可以偷懒：直接把完整句子当成一个事件（但结构化程度较低）
- 同样，也可以采用远程监督等弱监督的方式，实现限定域的事件抽取

- **开放域事件抽取**

- 限定域事件与预定义关系面临相似的问题：种类有限，维护困难。
- 如何在开放域环境下，自动识别未知结构与类型的事件？
  - 一种思路是采用无监督方法（从而摆脱对于标注语料的依赖），通过聚类找到潜在的事件簇
  - 该思路基于分布假设理论：
    - 候选事件触发词或者候选事件元素具有相似的语境，那么这些候选事件触发词倾向于触发相同类型的事件，相应的候选事件元素倾向于扮演相同的事件元素。
  - 然而，无监督事件抽取没有规范语义标签，难以映射到现有知识库。

- **开放域事件抽取**

- 另一种解决方法：与开放关系抽取中的“知识监督”方案类似
  - 主要挑战在于现有知识库中缺乏事件触发词信息，如何获取？
  - Y Chen, et al., Automatically Labeled Data Generation for Large Scale Event Extraction, ACL 2017
  - 定义事件核心元素，通过初步回标找到触发词并进行过滤和扩展



本章小结

知识图谱与图计算

- 图表示学习技术
  - 基于随机游走的图表示学习
  - 基于图神经网络的图表示学习
- 知识图谱表示学习
- 知识图谱推理补全
- 补充：事件抽取概述

tongxu@ustc.edu.cn