电子科技大学

University of Electronic Science and Technology of China

An Introduction to Knowledge Graph

Songling Liu

Data Mining Lab, Big Data Research Center, UESTC
Email： lsl571@163.com

➢Outline

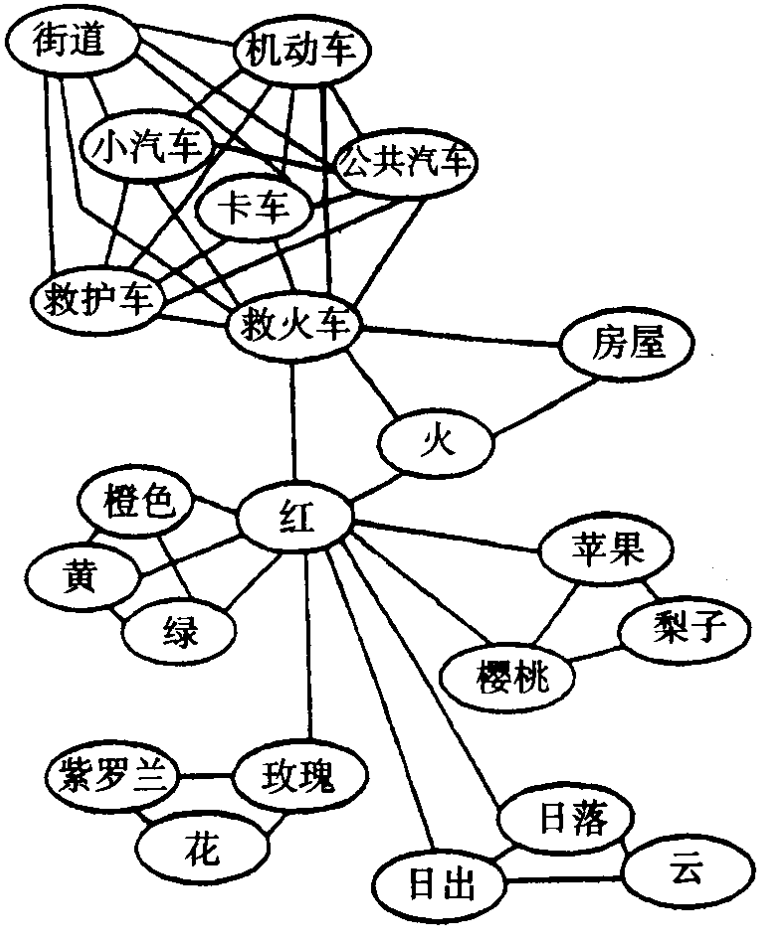
- 什么是知识图谱？
- 知识图谱的应用
- 知识图谱中的关键技术
  - 知识获取
  - 知识表示
  - 知识融合
  - 知识推理
- 总结

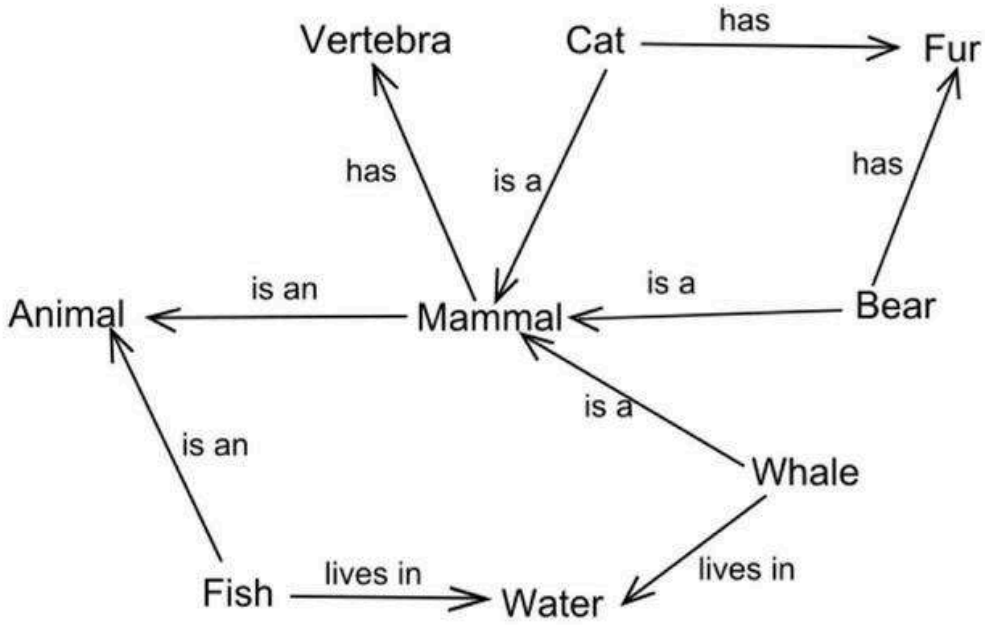
百度搜索


姚明的身高
网页 图片 新闻 视频 地图 更多 搜索工具
找到约 1,160,000 条结果 （用时 0.68 秒）
2.29 米
姚明，身高
沙奎尔·奥尼尔
2.16 米
勒布朗·詹姆斯
2.03 米
林书豪
1.91 米
反馈
姚明的身高的图片搜索结果
举报图片
姚明
篮球运动员
姚明，前中国篮球运动员，生于中国上海，祖籍为江苏苏州吴江区震泽镇，是原中国国家篮球队队员，曾效力于中国篮球职业联赛上海大鲨鱼篮球俱乐部和美国国家篮球协会休斯敦火箭。姚明是中国最具影响力的人物之一，同时也是世界最知名的华人运动员之一。2009年，姚明收购上海男篮，成为上海大鲨鱼篮球俱乐部老板。 维基百科
生于： 1980 年 9 月 12 日（34 岁），上海市
身高： 2.29 米
体重： 141 公斤
配偶： 叶莉（结婚时间：2007 年）
子女： 姚沁蕾
父母： 姚志源， 方凤娣

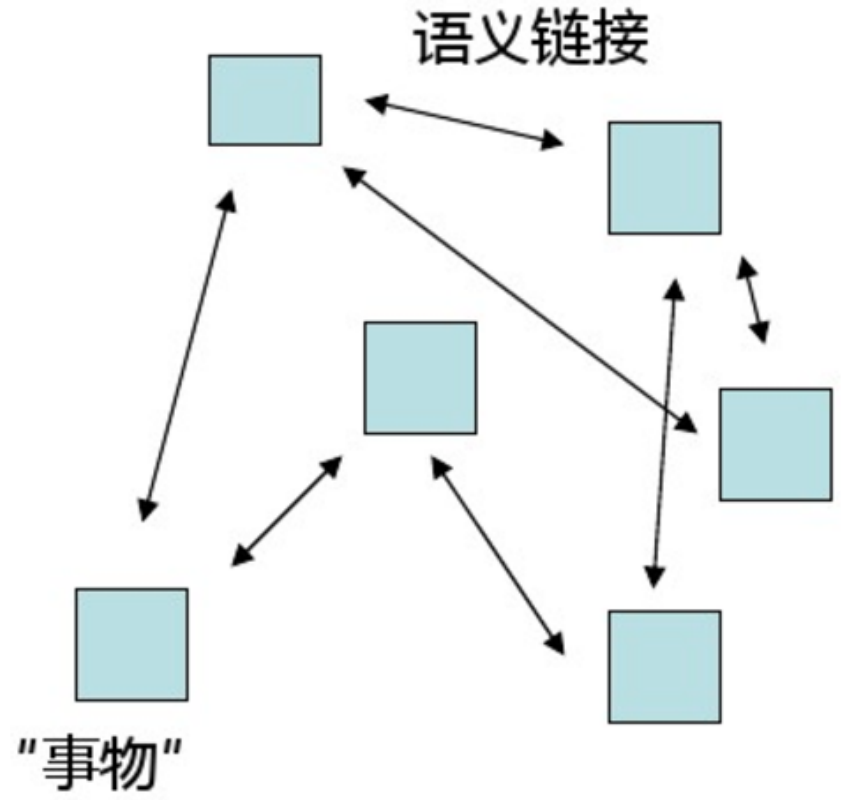

Bai du 百度
电子科技大学
百度一下
网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»
百度为您找到相关结果约15,900,000个
搜索工具
为您推荐: 西安电子科技大学 西南交通大学 重庆大学 电子科技大学信息门户
电子科技大学
信息公开 人才招聘 影像成电 电子科技博物馆 成电媒体 名师博客 中国政府网 教育部 工业和信息化部 四川省教育厅 中国大学生在线 清水河校区:成都市高新区(西区...
www.uestc.edu.cn/ - 百度快照 - 180条评价
电子科技大学_百度百科
电子科技大学（University of Electronic Science and Technology of China）简称"电子科大"，坐落于有"天府之国"之称的成都市，由中华人民共和国教育部直属，为教育部 、工信部 、四川省人民政府重点共建，位列"211工程"、"985工程...
成电校史 科学研究 教育教学 文化传统 学校领导 更多>>
baike.baidu.com/ -
相关院校
展开
清华大学
红色工程师的摇篮
两电一邮
指IT界中三所著名高校
成都信息工程大学
国家首批卓越
上海理工大学
信义勤爱 思学志远
四川理工学院
全日制公立高等院校
西南石油大学
明德笃志 博学创新
在成都市搜索电子科技大学_百度地图
郫都区
三河镇
安靖镇
金牛区
青羊区
成都市
A 电子科技大学(沙河校区)
5条评论
地址： 四川省成都市成华区建设北路二
B 电子科技大学(清水河校区)
26条评论
地址： 四川省成都市高新区(西区)西源
查看全部1192条结果>>
map.baidu.com

语义网络

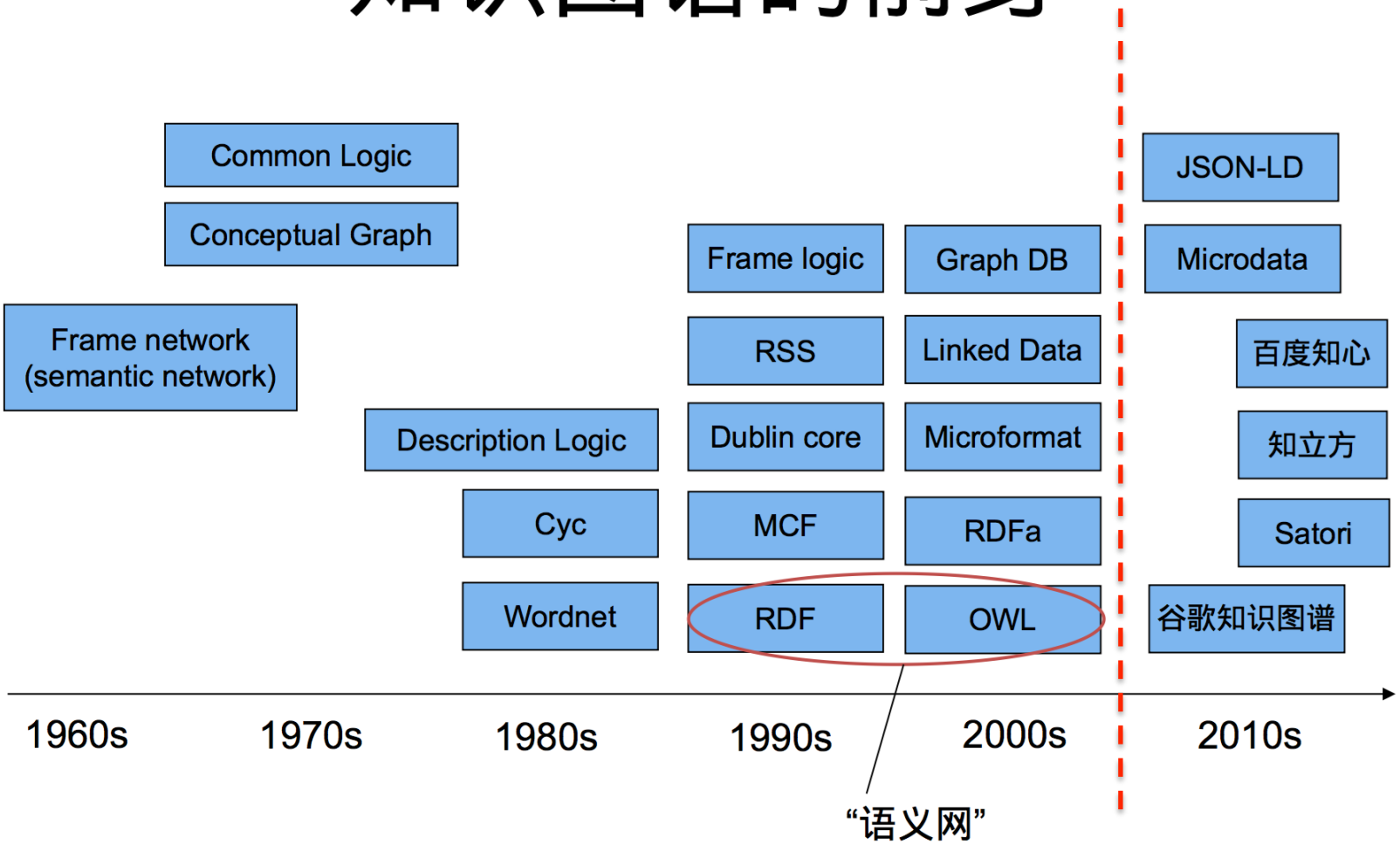

街道
机动车
小汽车
公共汽车
卡车
救护车
救火车
房屋
火
橙色
红
黄
绿
苹果
梨子
樱桃
紫罗兰
玫瑰
花
日落
日出
云


Vertebra
Cat
has
Fur
has
is a
has
is an
is a
Animal
Mammal
Bear
is an
is a
Whale
Fish
lives in
Water
lives in

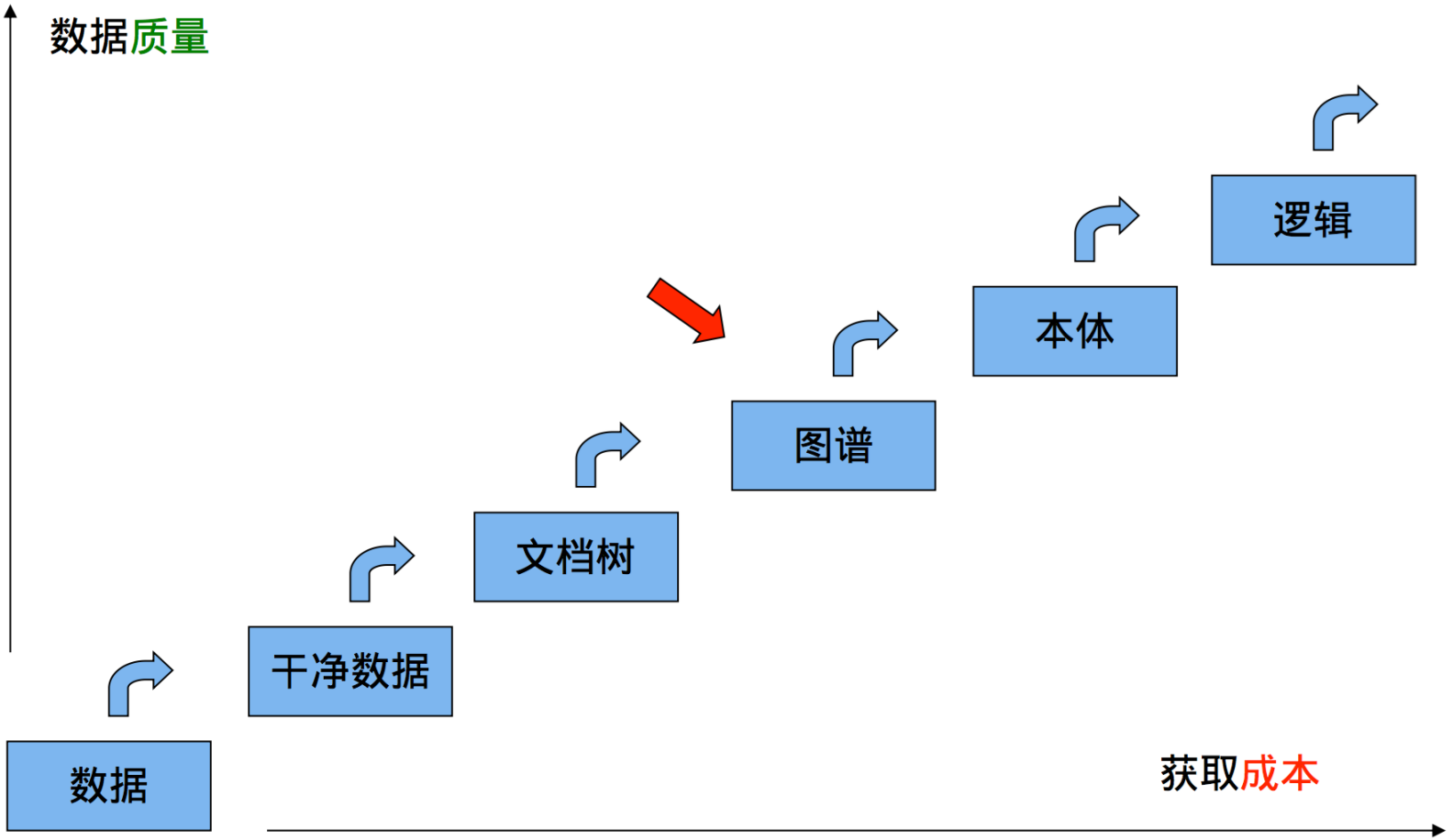
- 知识图谱：揭示实体之间关系的语义网络，对现实世界的事物及相互关系进行形式化地描述。（Things, not strings!）



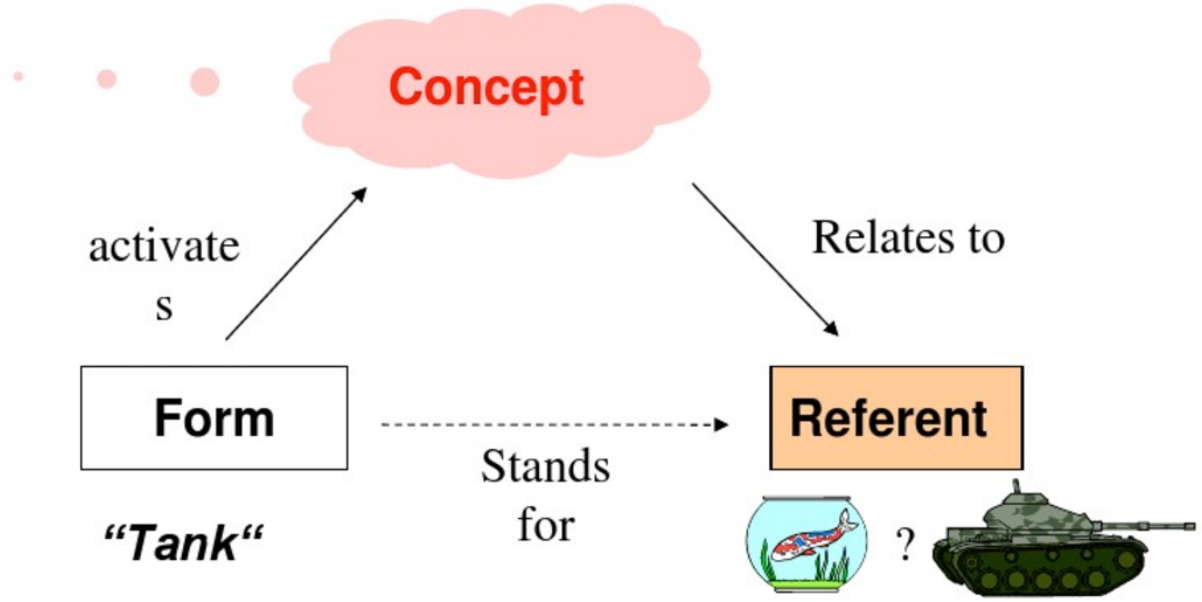
知识图谱的前身


Common Logic
Conceptual Graph
Frame network
(semantic network)
Description Logic
Cyc
Wordnet
Frame logic
RSS
Dublin core
MCF
RDF
Graph DB
Linked Data
Microformat
RDFa
OWL
JSON-LD
Microdata
百度知心
知立方
Satori
谷歌知识图谱
1960s
1970s
1980s
1990s
2000s
2010s
“语义网”

知识图谱



本体：



Ontology（中文称为：本体）是一种描述术语（包含哪些词汇）及术语间关系（描述苹果、香蕉、水果之间的关系）的概念模型。

本体的简化形式

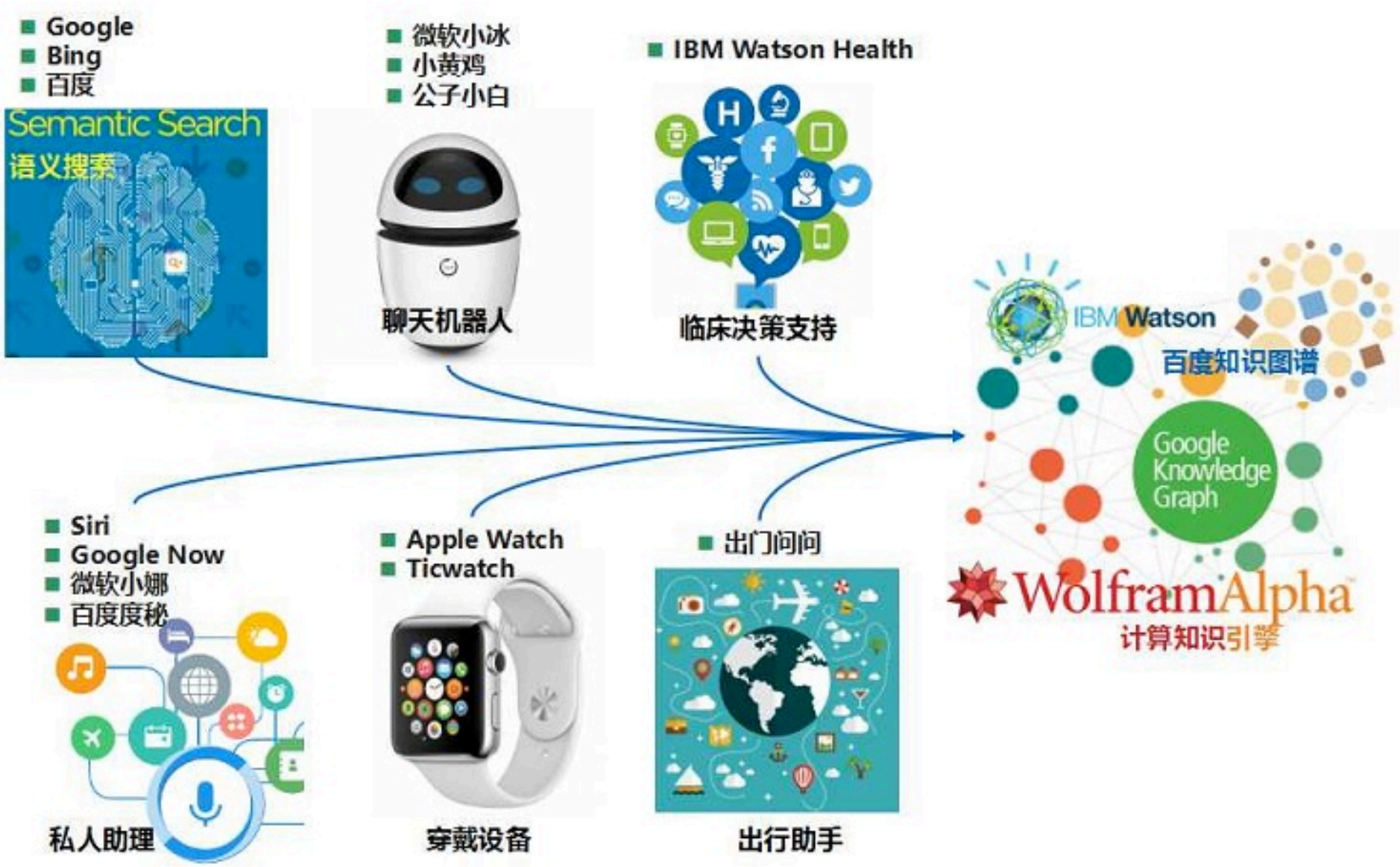
$$O = \{C, I, T, P\}$$

- *C – concepts*
  - 描述领域或任务中的**抽象概念**，通常以Taxonomy形式组织
  - 如描述世界知识的本体中，**学生**和**老师**是两个概念
- *I - instances*
  - 描述具体的**实例**
  - **学生Peter**是概念学生的实例
- *T - ISA*
  - **概念与概念之间、实例与概念之间的关系**
  - **subClassOf关系**和**instanceOf关系**
- *P – properties*
  - 本体中用于描述实例信息的**其他语义关系**
  - 如：**instance-attribute-value**（AVP）

➢知识图谱和本体的关系

- 知识图谱本身不是本体的一个替代品，是在本体的基础上面做了一个丰富和扩充，这种扩充主要体现在实体层面。
- 本体中突出的主要是概念和概念之间的关联关系，而知识图谱描述的主要是实体，对这些实体我们通常还会去描述它更加丰富的信息。
- 用一句简单的话来说就是：本体描述了知识图谱的数据模式，本体的动态的特性赋予了知识图谱动态数据模式支持的能力。

知识图谱的应用


Google
Bing
百度
Semantic Search
语义搜索
微软小冰
小黄鸡
公子小白
聊天机器人
IBM Watson Health
临床决策支持
IBM Watson
百度知识图谱
Google Knowledge Graph
WolframAlpha
计算知识引擎
Siri
Google Now
微软小娜
百度度秘
私人助理
Apple Watch
Ticwatch
穿戴设备
出门问问
出行助手

知识获取

知识表示

知识融合

知识推理

➢知识获取

知识抽取主要是面向开放的链接数据,通过自动化的技术抽取出可用的知识单元,知识单元主要包括实体(概念的外延)、关系以及属性3个知识要素, 并以此为基础,形成一系列高质量的事实表达。

- 知识提取是要解决结构化数据生成的问题。
- 但是广义上讲,知识提取是数据质量提升中的一环,各种提升数据质量的方法,都可以视为某种知识提取。
- 学术上一般是用自然语言处理的方法,但在实践中通常是利用规则。

实体提取

- 从百科类站点中提取
- 从垂直站点中提取
- 利用模式从网页和句子中提取
- 利用命名实体识别从自然语言句子中提取

表2　各种词表构建方法的比较（5表示在此指标上表现很好，1表示很差，“-”表示忽略该指标）

指标 / 方法	从百科类站点中提取	从垂直站点中提取	基于模式的方法	利用命名实体识别技术
精度	5	5	4	4
对领域和实体类型的覆盖率	5	2	5	2
实体覆盖率（在特定领域或类型上）	-	5	-	5
实体覆盖率（高频实体）	4	-	5	-
实体覆盖率（中低频实体）	2	-	4	-

开放类知识库

垂直行业知识库

➢命名实体识别（NER）

输入：无结构句子
输出：表明了命名实体类型的句子

1.基于规则与词典的方法：
需要限定文本领域、限定语义单元类型；人为定义规则，抽取出文本中的人名、地名、组织机构名、特定时间等实体；

缺点：基于规则模板的方法 不仅需要依靠大量的专家来编写规则或模板,覆盖的领域范围有限,而且很难适应数据变化的新需求。

2.基于统计机器学习的实体抽取方法：
对于每种实体类型提供少量种子实体，采用半监督的方法对实体进行提取；

缺点：需要针对每种类型提供种子实体或其他训练数据，难以扩展到开放域和所有类型对象。

➢知识表示

传统的表示形式即为三元组

$$\{Head, Relations, Tail\}$$

RDF:

- 资源—由唯一URI表示
- 属性—一种特殊的资源
- 语句—主语、谓语、宾语构成的三元组

但是其在计算效率、数据稀疏性等方面却面临着诸多问题。

表示学习的方法（向量表征）

- One-hot representation

star [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ⋯]
sun [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ⋯]

sim(star, sun) = 0 

- Distributed representation

单层神经网络模型, 对任意三元组{h,r,t}定义评价函数：

$$f_r(h,t) = \mu_r^T g(M_{r,1} l_h + M_{r,2} l_t)$$

Loss function：

$$J(\mathbf{\Omega}) = \sum_{i=1}^{N} \sum_{c=1}^{C} \max\left(0, 1 - g\left(T^{(i)}\right) + g\left(T_c^{(i)}\right)\right) + \lambda \|\mathbf{\Omega}\|_2^2,$$

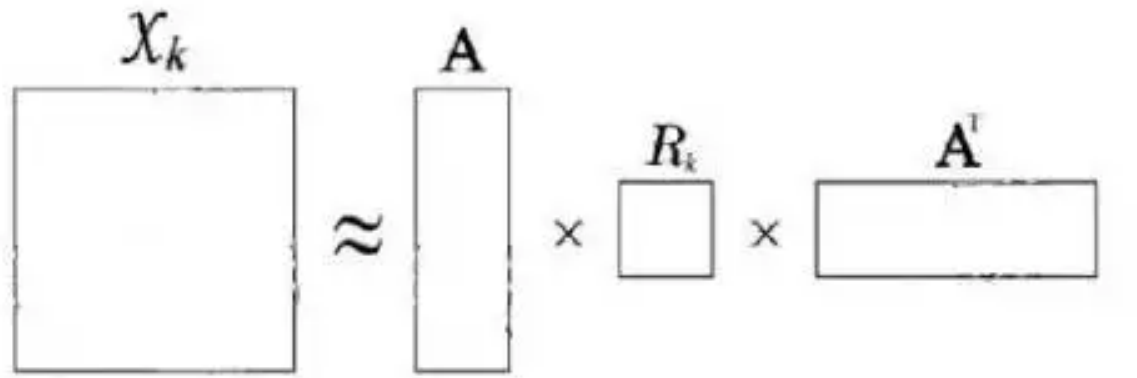
$$T_c^{(i)} = (e_1^{(i)}, R^{(i)}, e_c).$$

双线性模型, 对任意三元组{h,r,t}定义评价函数：

$$f_r(h,t) = l_h^T M_r l_t$$

M_r 是关系r对应的双线性变换矩阵，通过简单有效的方法刻画了实体与实体之间的语义关系，协同性好，计算复杂度低。

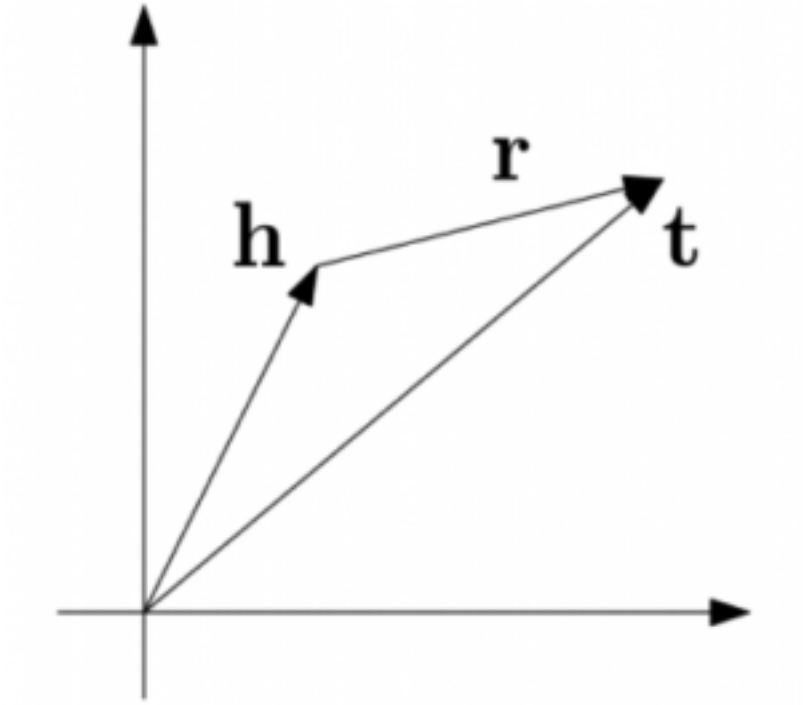
矩阵分解模型，知识库中的三元组集合表示为一个三阶张量



$$\min_{A,\{R_k\}} \sum_{k=1}^{K} || \chi_k - AR_k A^T ||_F^2 + \lambda_1 || A ||_F^2 + \lambda_2 \sum_{k=1}^{K} || R_k ||_F^2$$

翻译模型，受到平移不变现象的启发,提出了TransE模型，对于任意三元组{h,r,t}

$$l_h + l_r \approx l_t$$



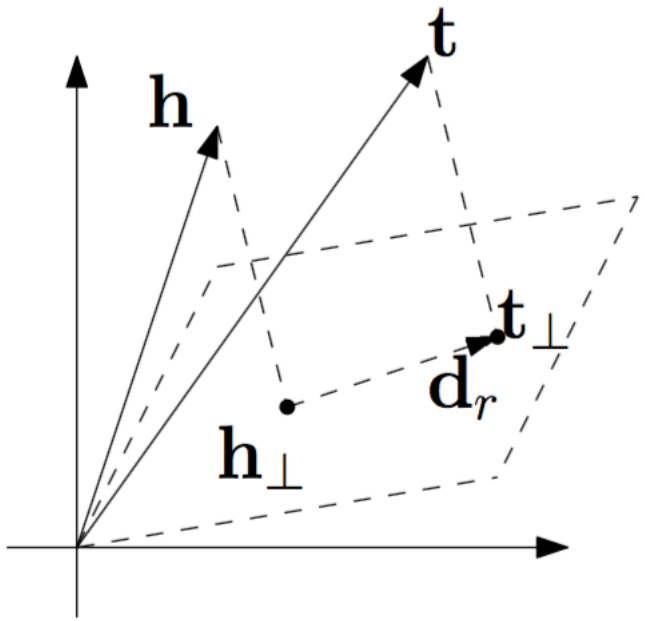
Loss function

$$f_r(h,t) = ||l_h + l_r - l_t||_{L_1/L_2}$$

TransH

TransH模型尝试通过不同的形 式表示不同关系中的实体结构,对于同一个实体而 言,它在不同的关系下也扮演着不同的角色；

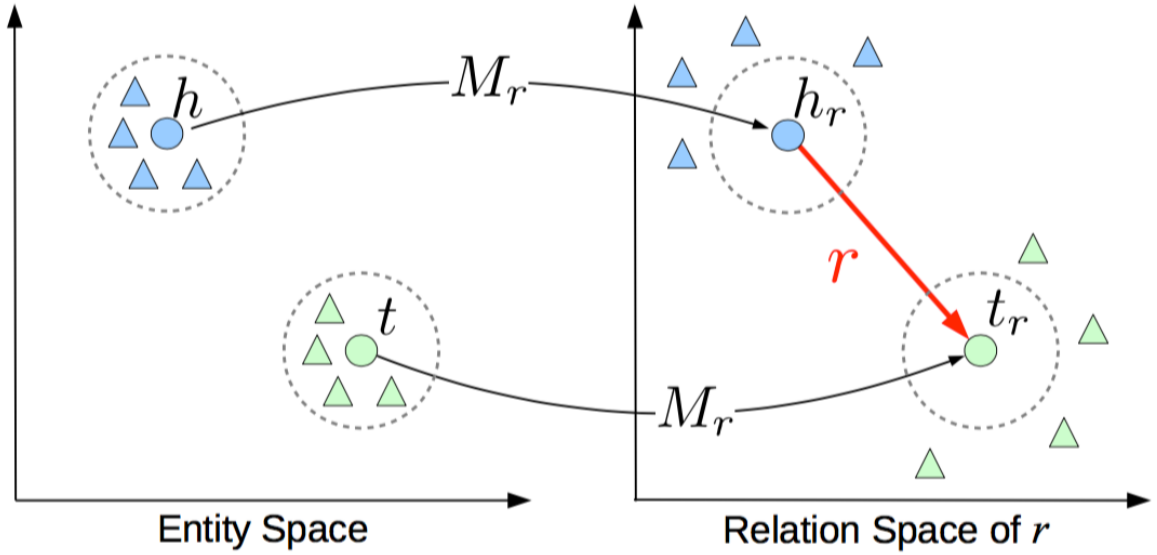
$$h_{\perp} + r \approx t_{\perp}$$



但是该假设中，实体和关系还是存在于相同的语义空间当中。

TransR

将知识库中的每个三元组(*h*,*r*,*t*)的头实体与尾实体向关系空间中投影



$$h_r = hM_r$$

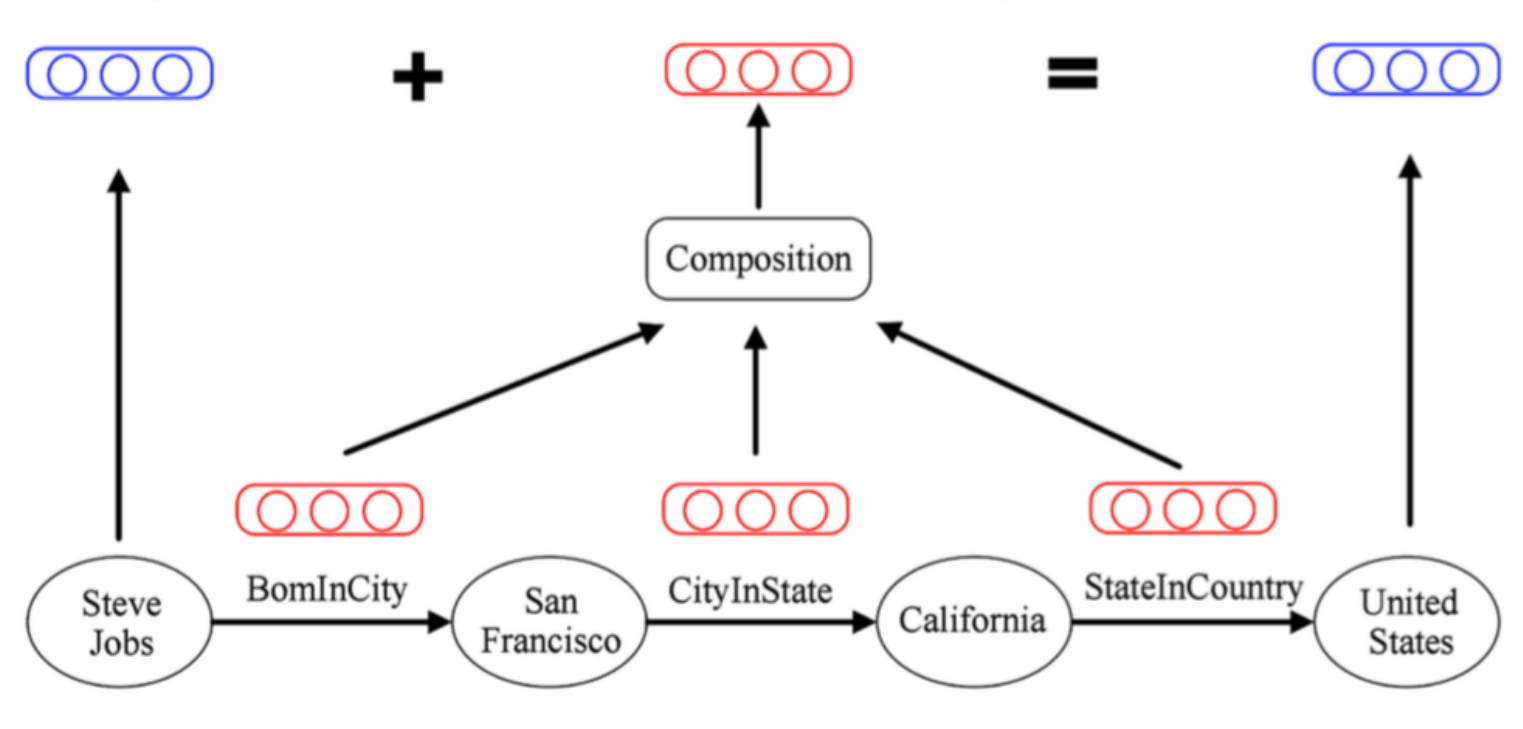
$$t_r = tM_r$$

Loss function

$$f_r(h,t) = \| h_r + r - t_r \|_2^2$$

PTransE

涉及实体多步连接的关系序列，考虑到多步关系能提高表示学习的区分性，在知识补全任务中发挥更好的作用。



➢知识融合

由于知识图谱中的知识来源广泛,存在知识质量良莠不齐、来自不同数据源的知识重复、知识的关联不够明确等问题,所以必须要进行知识的融合。

- 实体对齐： 主要是用于消除异构数据中实体冲突、指向不明等不一致性问题,可以从顶层创建一个大规模的统一知识库,从而帮助机器理解多源异质的数据,形成高质量的知识。

- 本体构建： 本体是同一领域内不同主体之间进行交流、连通的语义基础，有利于进行约束、推理等,却不利于表达概念的多样性。

- 质量评估 ： 对知识库的质量评估任务通常是与实体对齐任务一起进行的,其意义在于,可以对知识的可信度进行量化,保留置信度较高的,舍弃置信度较低的, 有效确保知识的质量。

➢知识推理

根据知识图谱中已有的知识推断出新的、未知的知识。知识推理的对象可以是实体、实体的属性、实体间的关系、本体库中概念的层次结构等。

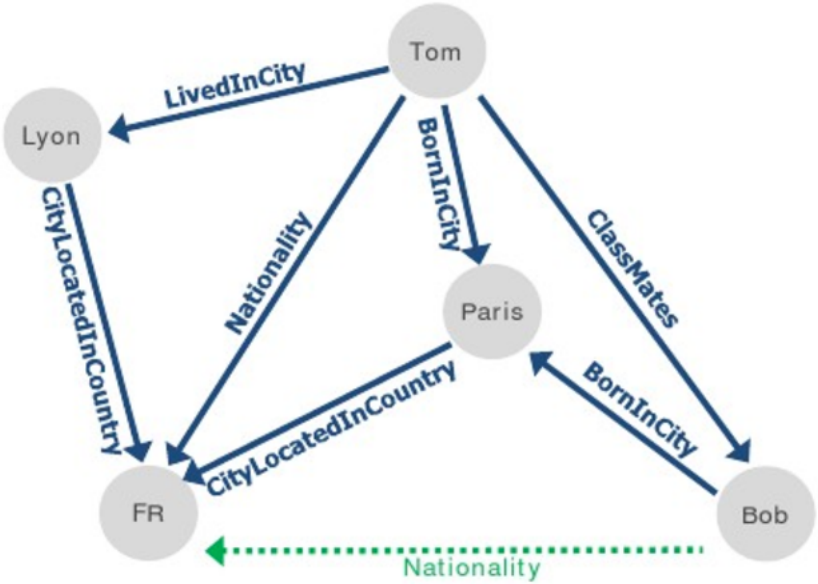
✦ 基于逻辑的推理：基于逻辑的推理方式主要包括一阶谓词逻辑 (first order logic)、描述逻辑(description logic)以及规则等。

✦ 基于图的推理：用于推理实体间隐含的关系。主要是利用了关系路径中的蕴涵信息,通过图中两个实体间的多步路径来预测它们之间的语义关系。

- Path Ranking 算法（PRA），通过连接实体的已有路径来预测实体间的潜在关系；
- 基于表示学习的模型，将实体和关系映射为空间中的向量，通过空间中向量的运算来进行推理（如TransE）；
- 概率图模型，如马尔科夫逻辑网络及其衍生物。

PRA模型

PRA方法具有较好的解释性，并且不需要额外的逻辑规则。在利用PRA进行关系推理时，利用PRA为每个关系独立建模，也就是为每个关系学习一个独立的分类器。

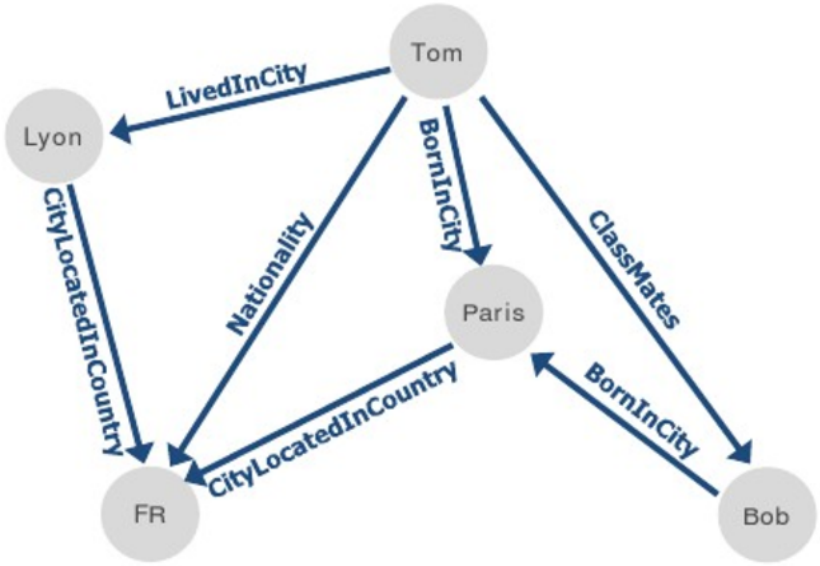


(Tom, BornInCity, Paris)
(Tom, LivedInCity, Lyon)
(Tom, Nationality, France)
(Tom, ClassMates, Bob)
(Paris, CityLocatedInCountry, France)
(Lyon, CityLocatedInCountry, France)
(Bob, BornInCity, Paris)
(Bob, Nationality, France)

PRA模型

核心思想：以路径作为特征训练关系专属分类器

-- 路径：链接两个实体的关系序列



Target Relation

bornInCity

Positive Instances

(Tom, Paris), (Bob, Paris)

Negative Instances

(Tom, Lyon), (Bob, Lyon)

Feature Set

$\text{nationality} \rightarrow \text{cityLocatedInCountry}^{-1}$

$\text{classMates} \rightarrow \text{bornInCity}$

$\text{classMates}^{-1} \rightarrow \text{bornInCity}$

$\text{classMates}^{-1} \rightarrow \text{livedInCity}$

Training Instances

$\{(1,1,0,0),1\}$, $\{(0,0,1,0),1\}$

$\{(1,0,0,0),-1\}$, $\{(0,0,0,1),-1\}$

PRA模型的基本流程：

- **特征抽取**
  - -- 随机游走，广度优先搜索，深度优先搜索
- **特征计算**
  - -- 随机游走概率，布尔值，出现频次
- **分类器训练**
  - -- 单任务学习：为每个关系单独训练一个二分类分类器
  - -- 对任务学习：将不同关系进行联合学习，同时训练他们的分类器

➢Outline

- 什么是知识图谱？
- 知识图谱的应用
- 知识图谱中的关键技术
  - 知识获取
  - 知识表示
  - 知识融合
  - 知识推理
- 总结

Thanks