北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

大数据分析与应用-知识图谱

Big Data Analysis and Application - Knowledge Graph

汇报人：郭沛祺 王余阳 朱逸铭
张羽冰 杨笔奇 杨松坤

时间：2021.11.15

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目 录

CONTENTS

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

概述

Summary

汇报人：杨松坤

●○○○○

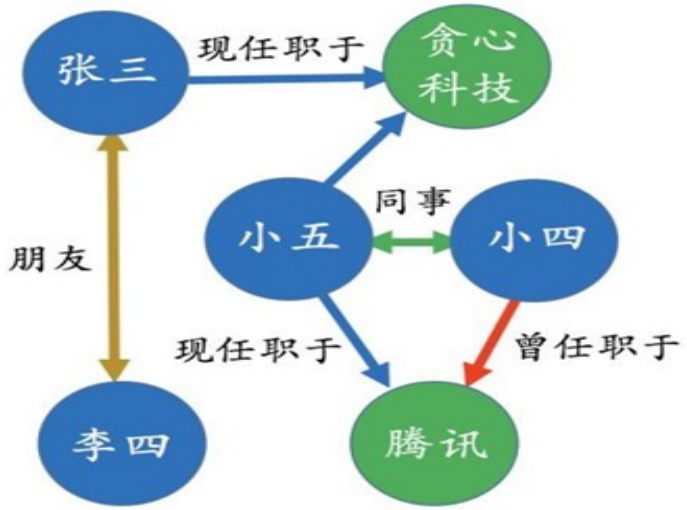
Knowledge Graph

知识图谱（Knowledge Graph）是google于2012年提出的概念，本质是语义网络（Semantic Network）的知识库，也可理解为多关系图，是能理解真实世界实际事物关联（real-world entities connections）的智能模型。

本质-多关系图

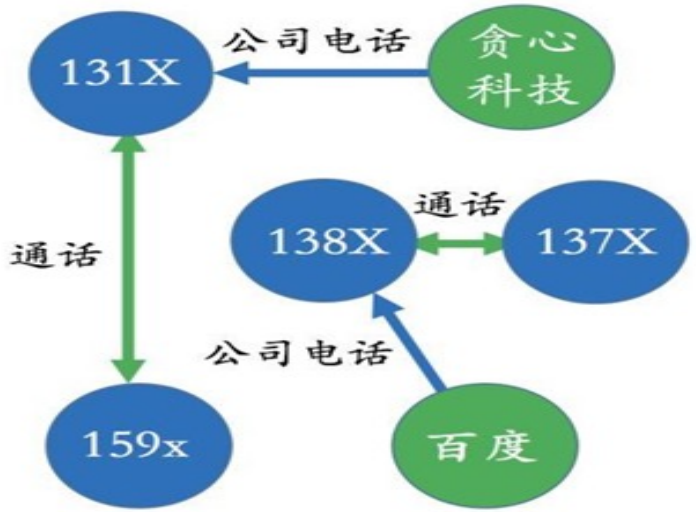
多关系图是包含多种类型节点与边的图。实体（节点）是真实世界中事物的**抽象**，关系用来刻画实体间的**联系**。

中国古人的五行学说就是对世间万物进行高度抽象后的**知识图谱**。



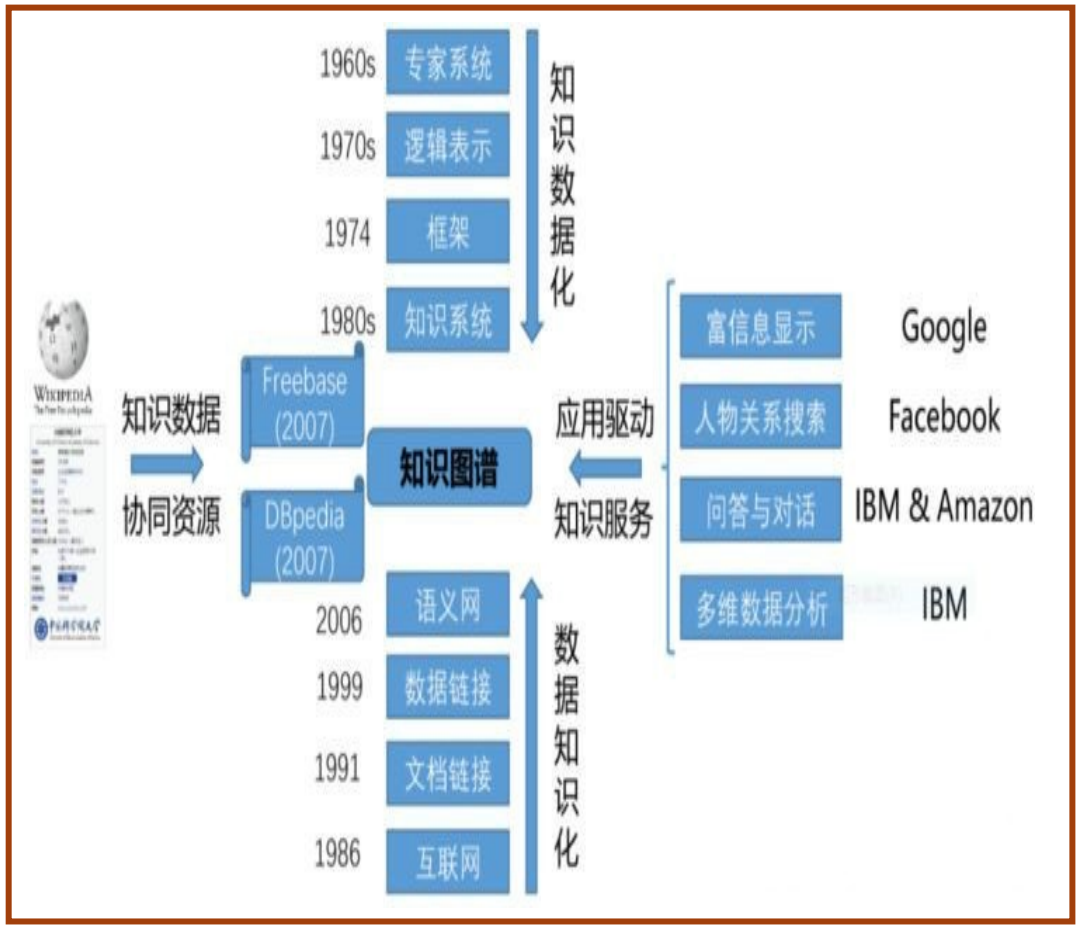
案例：社交网络

知识图谱表达现实世界场景



案例：风控知识图谱

比如一个社交网络图谱里，我们既可以有“人”和“公司”作为**实体**。人和人之间的**关系**可以是“朋友”，也可以是“同事”。人和公司之间的**关系**可以是“现任职”或者“曾任职”；类似的，一个风控知识图谱可以包含“电话”、“公司”的**实体**，电话和电话之间的**关系**可以是“通话”，而且每个公司它也会有固定的电话。



知识图谱发展历程

知识的数据化

通过将知识用计算机进行表示和组织，并设计相应算法完成推理、预测等任务。**专家系统**就是利用知识库支撑AI的一种有效尝试。

数据的知识化

通过引入知识，使得原始数据能够支撑推理、问题求解等复杂任务。这个目标的实践者就是**语义网**。

知识图谱按领域可分为通用知识图谱和行业知识图谱。

通用知识图谱

- 注重广度，强调融合更多的实体
- 主要应用于智能搜索等领域
- 准确度不够高，并且受概念范围的影响

行业知识图谱

- 依靠特定行业的数据来构建，具有特定的行业意义
- 实体的属性与数据模式比较丰富，需要考虑不同的业务场景与使用人员

知识图谱框架

逻辑架构

数据层

由一系列的事实组成，知识以事实为单位存储在图数据库

模式层

构建在数据层之上，是知识图谱的核心，存储经过提炼的知识

技术架构

自顶向下构建

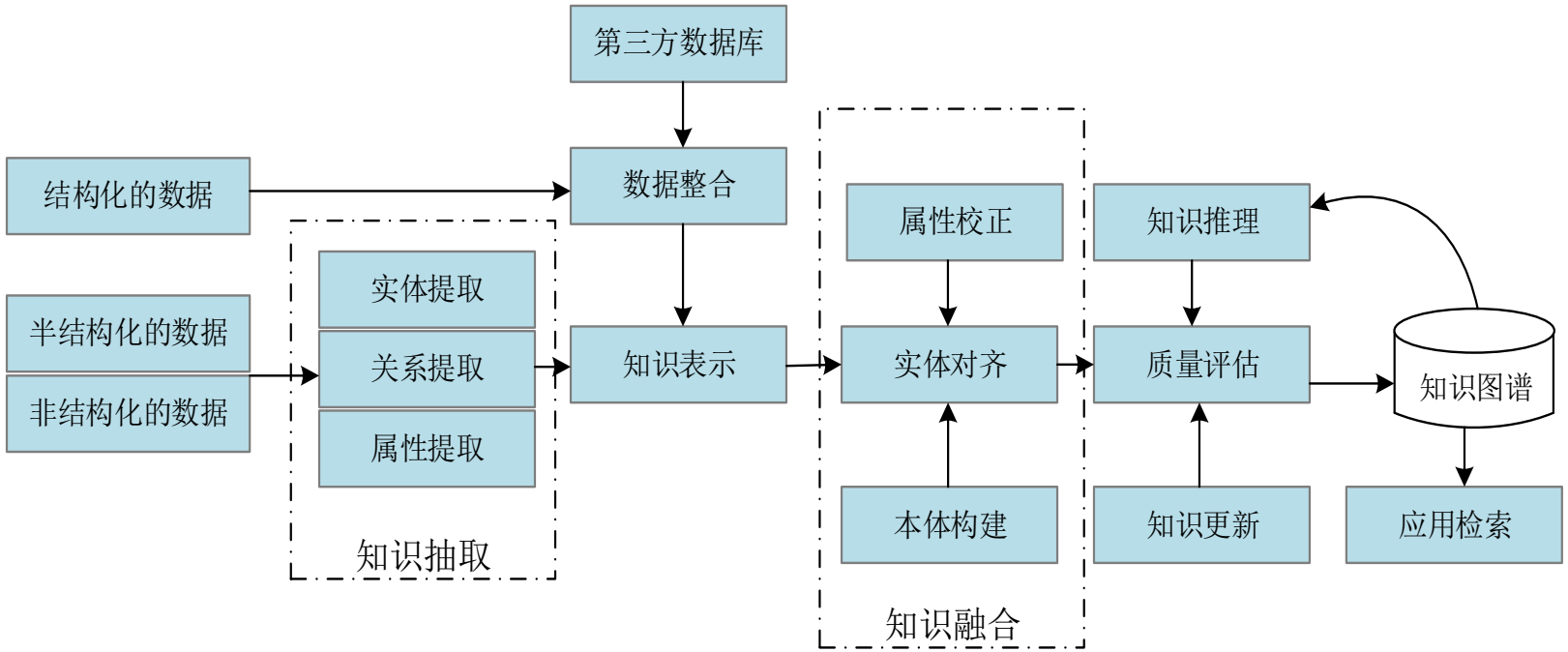
借助百科类网站等结构化数据源从高质量数据中提取本体和模式信息

自底向上构建

借助技术手段，从公开采集的数据中提取出资源模式

知识图谱构建过程

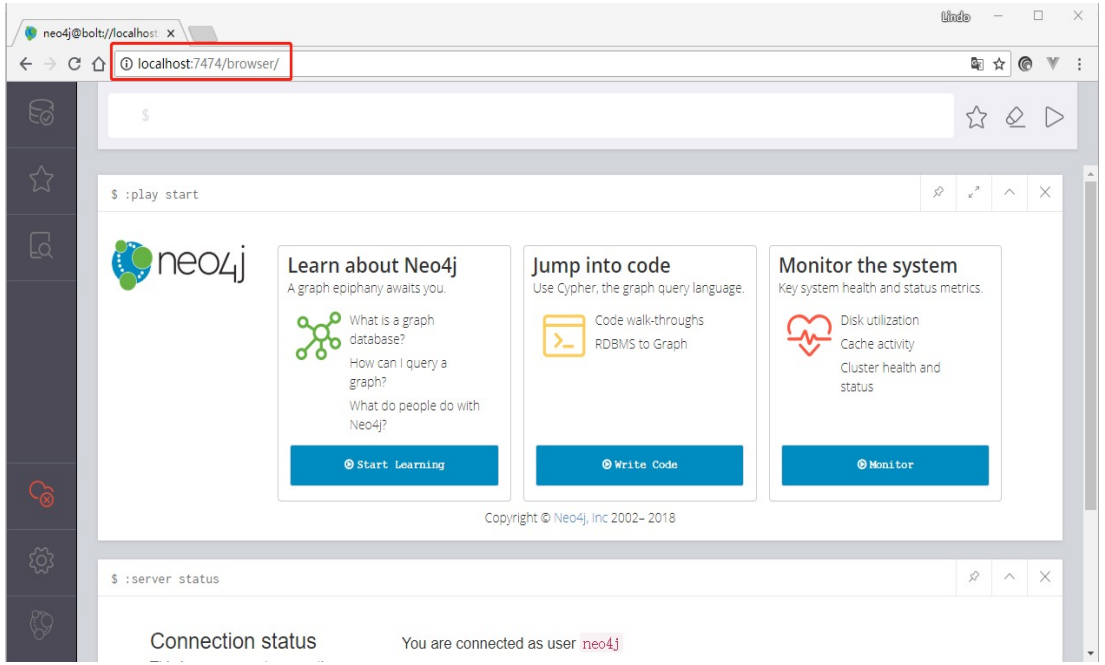
是从原始数据出发，采用一系列自动或半自动的技术手段，从原始数据中提取出知识要素（即事实），并将其存入知识库的数据层和模式层的过程。这是一个迭代更新的过程，根据知识获取的逻辑，每一轮迭代包含4个阶段：知识抽取、知识表示、知识融合以及知识推理。



知识图谱存储

常见的关系型数据库诸如MySQL之类不能很好的体现知识图谱数据的实体、属性、关系等。

同时，知识的组织形式采用的就是图结构，因此存储采用图数据库，以Neo4j最为常见。



neo4j存储优势

Neo4J属于原生图数据库，在图上互相关联的节点在数据库中的物理地址也指向彼此，因此更能发挥出图结构形式数据的优势。性能上对长程关系的查询速度快，也擅于发现隐藏的关系。

neo4j存储形式

主要是node和edge来组织数据。node可以代表实体，edge代表实体间的关系，关系可以有方向。另外，可以在node上加一个或多个标签表示实体的分类，以及一个键值对集合来表示该实体除了关系属性之外的一些额外属性。关系也可以附带额外的属性。

知识图谱的作用价值：

助力搜索

一方面通过推理实现概念检索，另一方面以图形化方式展示经过分类整理的结构化知识

助力问答

问答与对话系统一直是NLP在人工智能实现领域的关键标志之一。知识图谱给问答与对话系统挂载了一个背景知识库。

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

应用

Application

汇报人：王余阳

常见知识图谱

偏向于语言学的WordNet(主要用以英语的词义消歧)

偏向于概念层面的ConceptNet

常识知识库：Web Child、Cyc

领域知识库：HerbNet(中草药)，linked life data(生命科学领域)

单语言或者多语言的百科数据：YaGo，Freebase，Zhishi.me，复旦CN-DBpedia，Wikidata，PKUBase，清华XLORE

典型的中文知识图谱

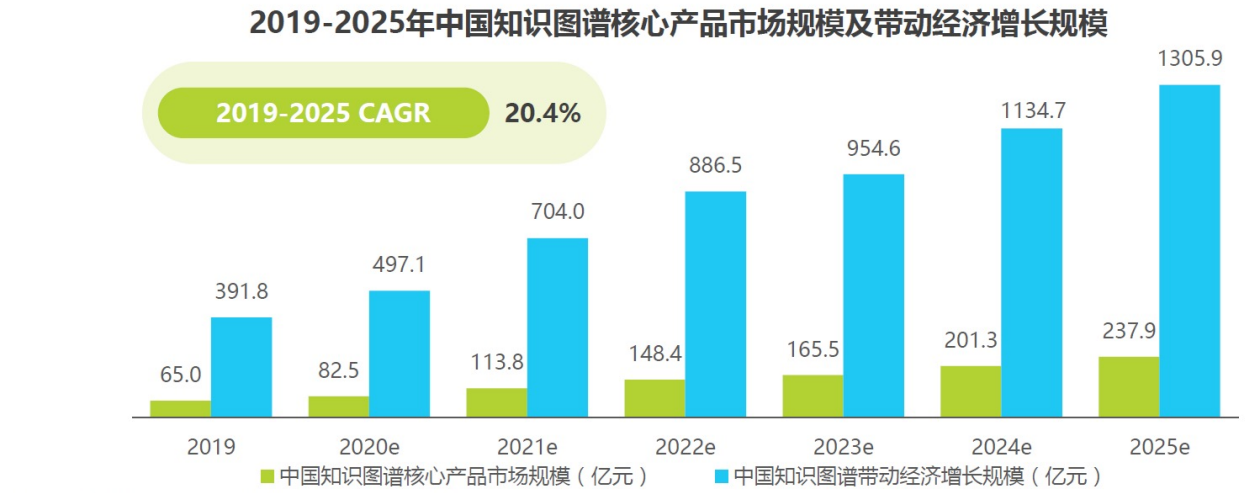
Zhishi.me	XLORE
针对单数据源深入挖掘，经知识抽取、知识清洗、知识填充以及知识更新等操作后，最终形成一个质量高、知识多、更新快的中文通用百科知识图谱，它现在包含超过 2000 万个实例。	融合中英文维基、法语维基和百度百科，对百科知识进行结构化和跨语言链接构建的多语言知识图谱，是中英文知识规模较平衡的大规模多语言知识图谱，它现在包含超过 1600 万个实例。
特点：拥有实例多，查询结果较为较为正确和全面； 提供结构化数据；	特点：具有更丰富的语义关系，基于isA关系验证； 拥有多种查询接口，助力第三方使用；
不足：匹配其它语言准确性低；	不足：英文实例和中文实例之间的跨语言链接数量有限； 实例的类型信息不完整。

知识图谱产业链

当前已初步呈现出知识图谱供应商、集成商、用户企业及基础工具服务商等生态合作伙伴协同发展的产业生态体系框架。

中国知识图谱市场规模

据艾瑞咨询统计推算，2019年知识图谱核心产品的市场规模约为65.0亿元，预计2024年将突破200亿元，年复合增长率达到20.4%。



来源：艾瑞咨询研究院根据专家访谈、招投标项目统计推算。
注释：知识图谱核心产品：Schema三元组模型构建、实体标注等技术，知识图谱管理平台与建模服务、垂直行业的知识图谱应用产品及解决方案；知识图谱带动收入：带动大数据、智慧应用以及传统产业效益提升规模。

知识图谱在各领域中的应用概览

从右图可以看出来，数据繁杂、单一价值有限、问题抽象需要可视化展现、五层关联维度以上的应用场景更加适合搭建知识图谱

	行业知识库	关联搜索	预警应用	研判应用	推荐应用	数据中台
金融领域	✓	✓	✓	✓	✓	✓
公安领域	✓	✓	✓	✓		
医疗领域	✓	✓	✓	✓	✓	
教育领域	✓	✓		✓	✓	✓
能源领域	✓	✓	✓	✓		✓
工业领域	✓	✓	✓	✓		✓
司法领域	✓	✓		✓	✓	
零售电商领域	✓	✓	✓		✓	✓
政务领域	✓	✓	✓	✓	✓	✓
客服领域	✓	✓			✓	
营销领域	✓	✓		✓	✓	
媒体舆情领域	✓	✓	✓	✓		
企服领域	✓	✓			✓	✓

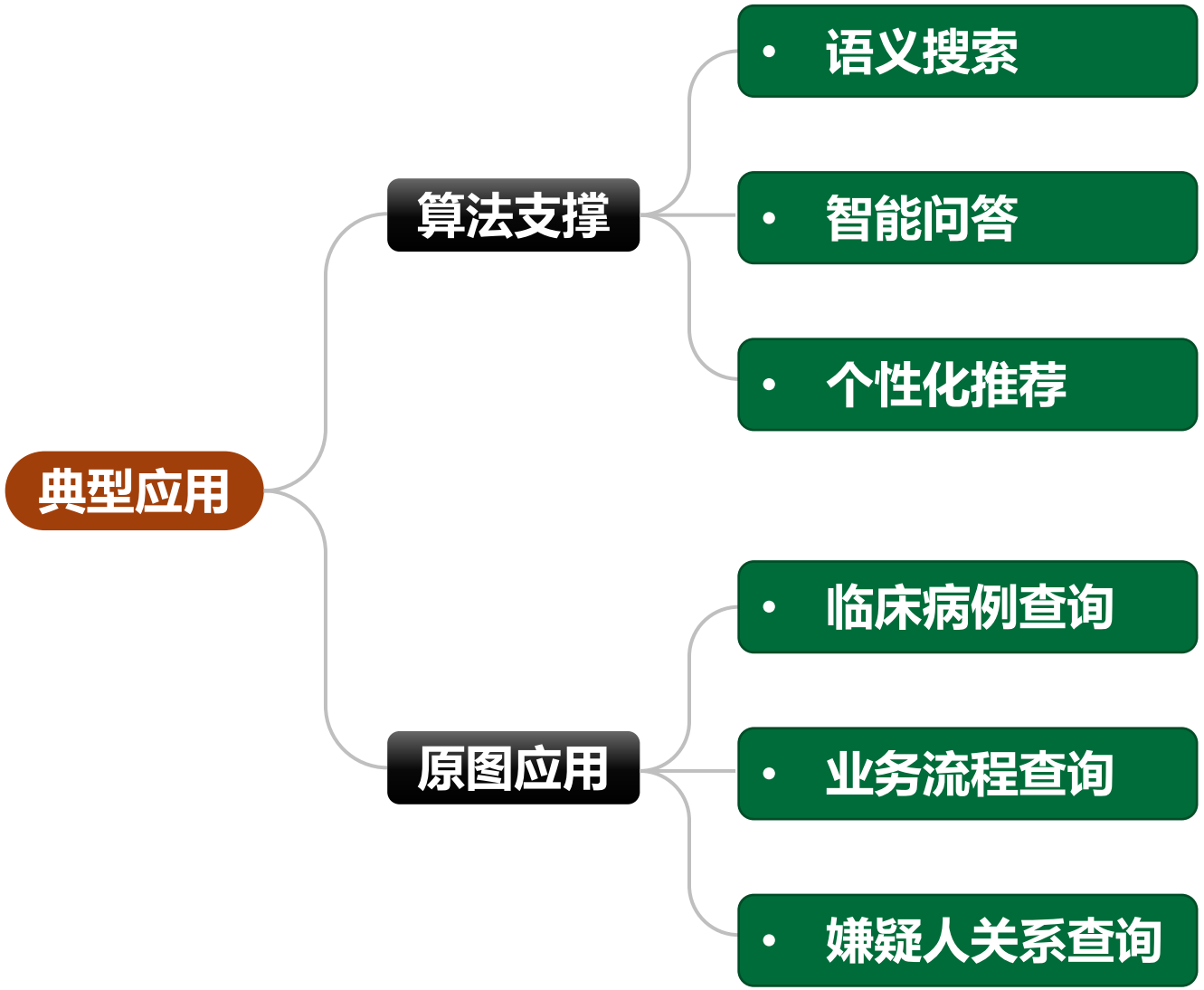
知识图谱应用

算法支撑：指通过知识图谱对于信息源的数据进行处理，将产出的结构化关联数据用于其他人工智能任务的算法模型训练和应用中，得到能解决具体场景问题的研判建议，形成解决办法产生价值的服务形式。

知识图谱应用类别：

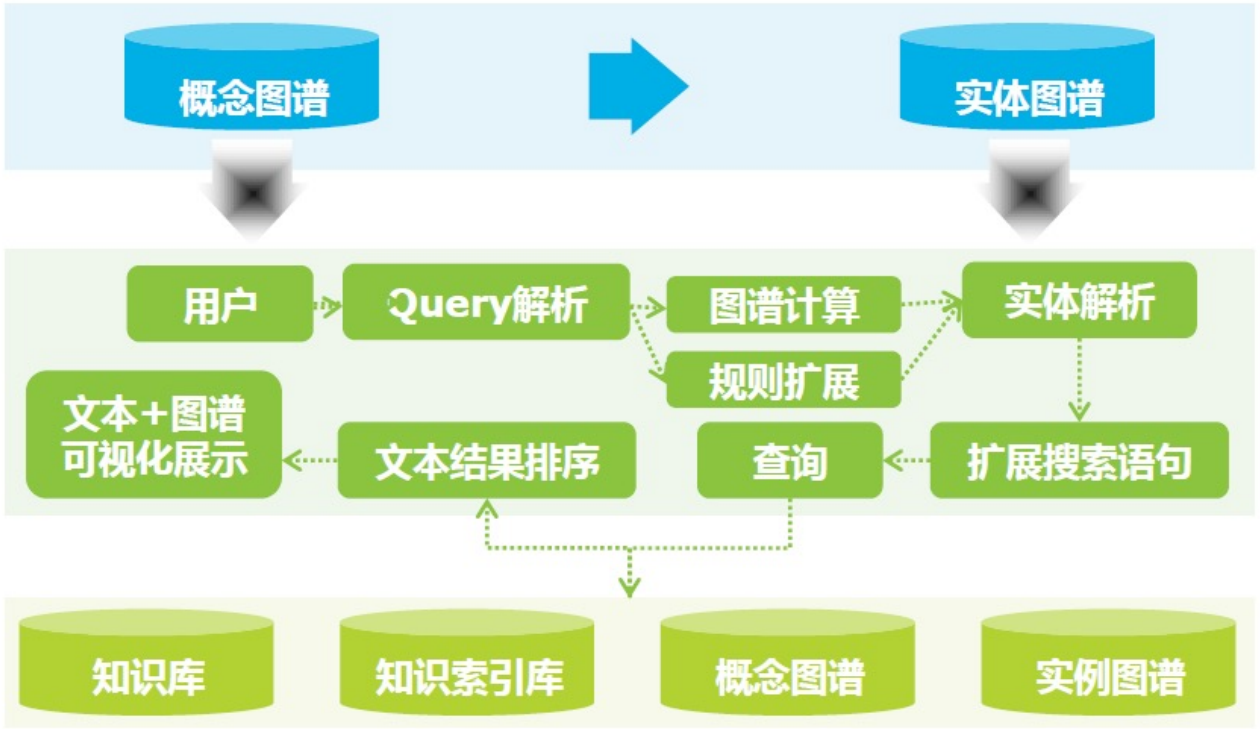
- 算法支撑
- 原图应用

原图应用：指基于知识图谱的图结构和丰富的语义关系，直接通过图谱产生价值的服务形式，例如图挖掘、关联分析等。



举例说明 — 语义搜索

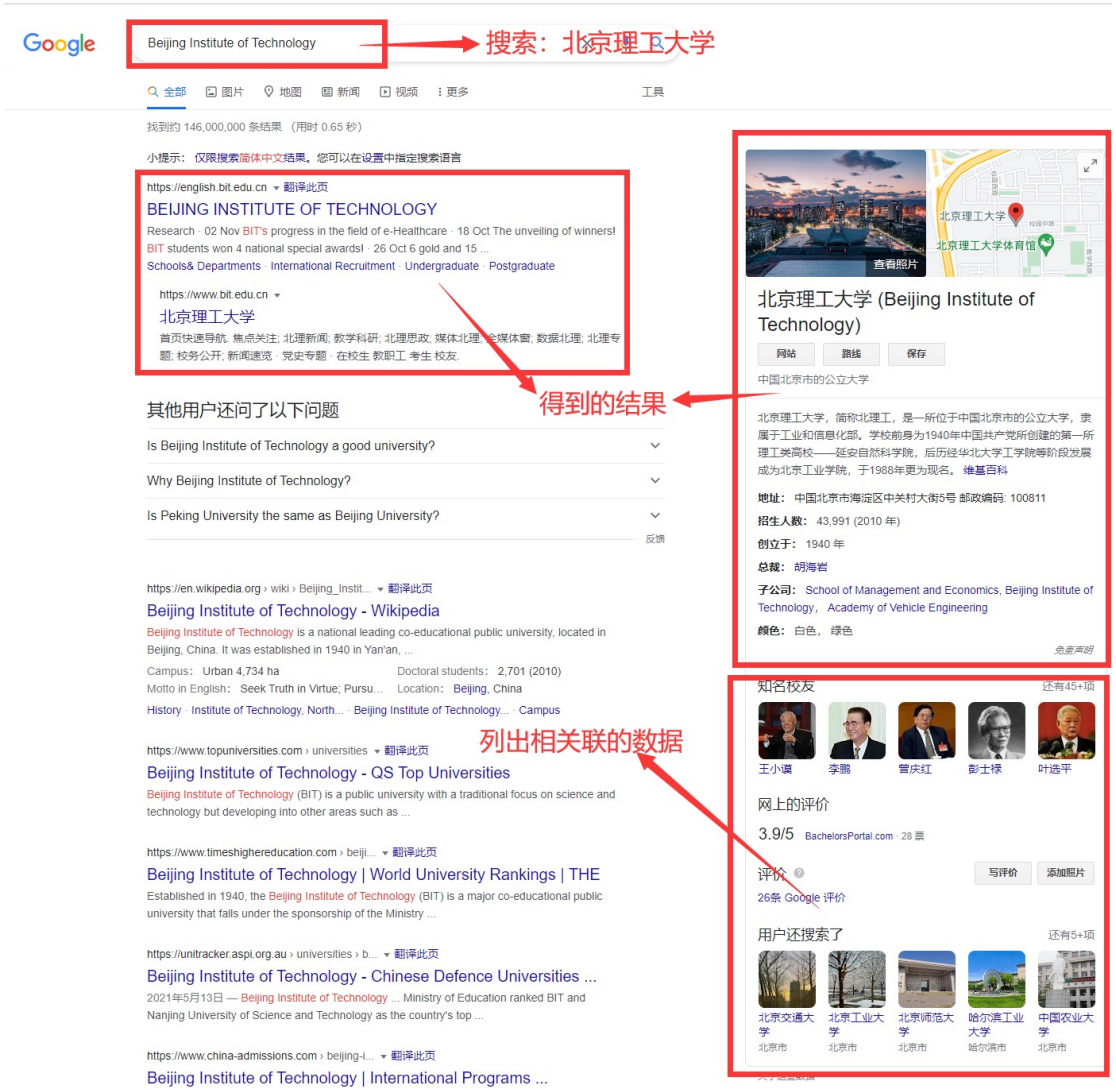
指搜索引擎的工作不再拘泥于用户所输入请求语句的字面本身，而是透过现象看本质，准确地捕捉到用户所输入语句后面的真正意图，并以此来进行搜索，从而更准确地向用户返回最符合其需求的搜索结果。



知识图谱应用

这里以google浏览器搜索“北京理工大学”为例：

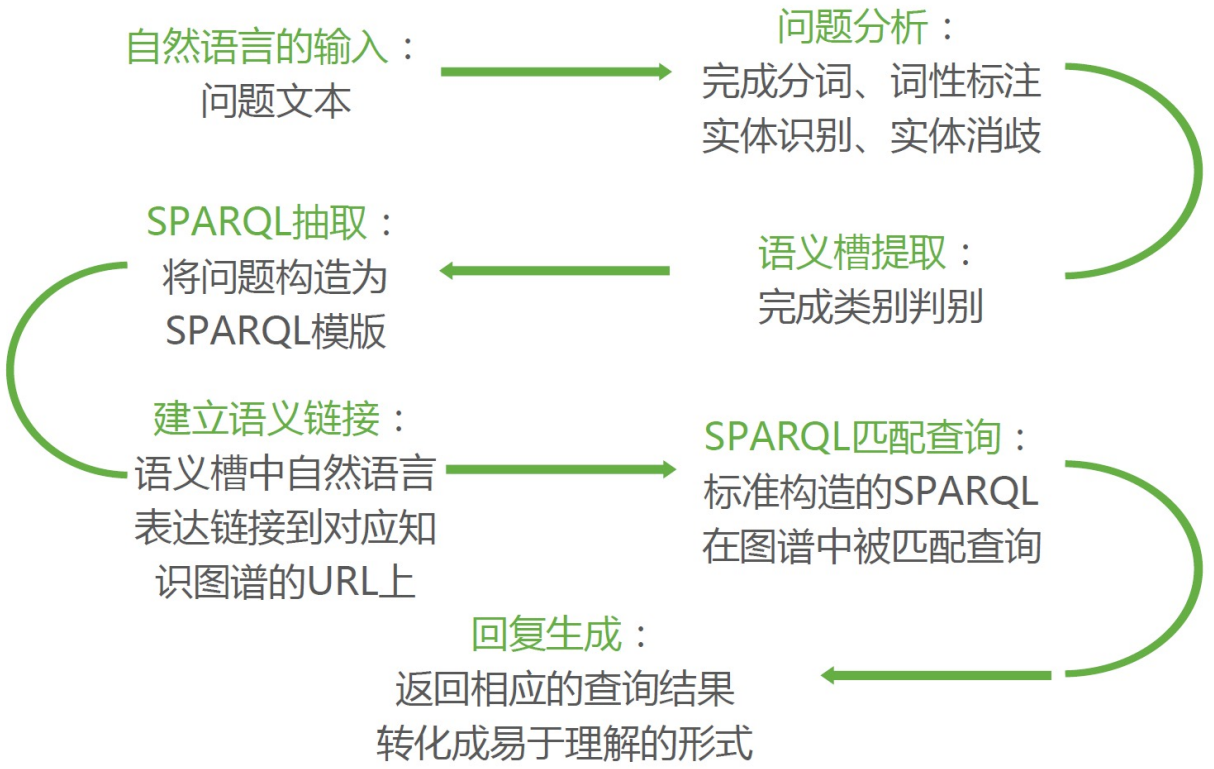
语义搜索	传统搜索
搜索引擎识别出“Beijing Institute of Technology”是“北京理工大学”，而且会给出京理工大学的各种属性信息，比如说地址、创建时间、招生人数等，这些都是以前基于关键词的检索做不到的，有了知识图谱以后，就可以即问即答了。	Google 给你呈现的是一页包含这些关键词的链接，Google 并不知道这个问题的真正含义。



举例说明 — 智能问答

智能问答以一问一答形式，精确的定位网站用户所需要的提问知识，通过与网站用户进行交互，为网站用户提供个性化的信息服务。

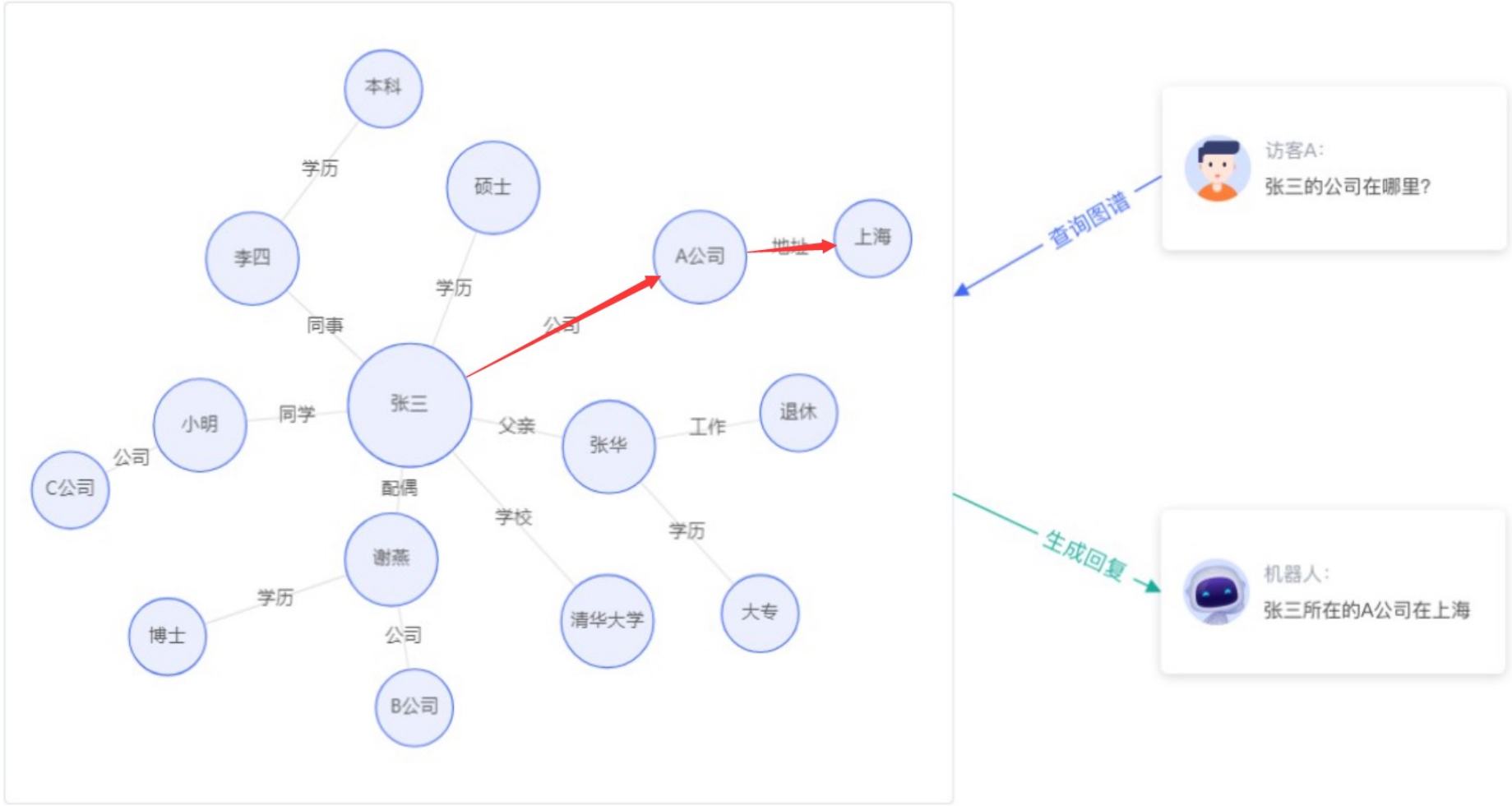
我们熟知的应用有各大手机平台里面的语音助手，例如Siri,，还有智能机器人等。



知识图谱应用

以智能机器人为实例：

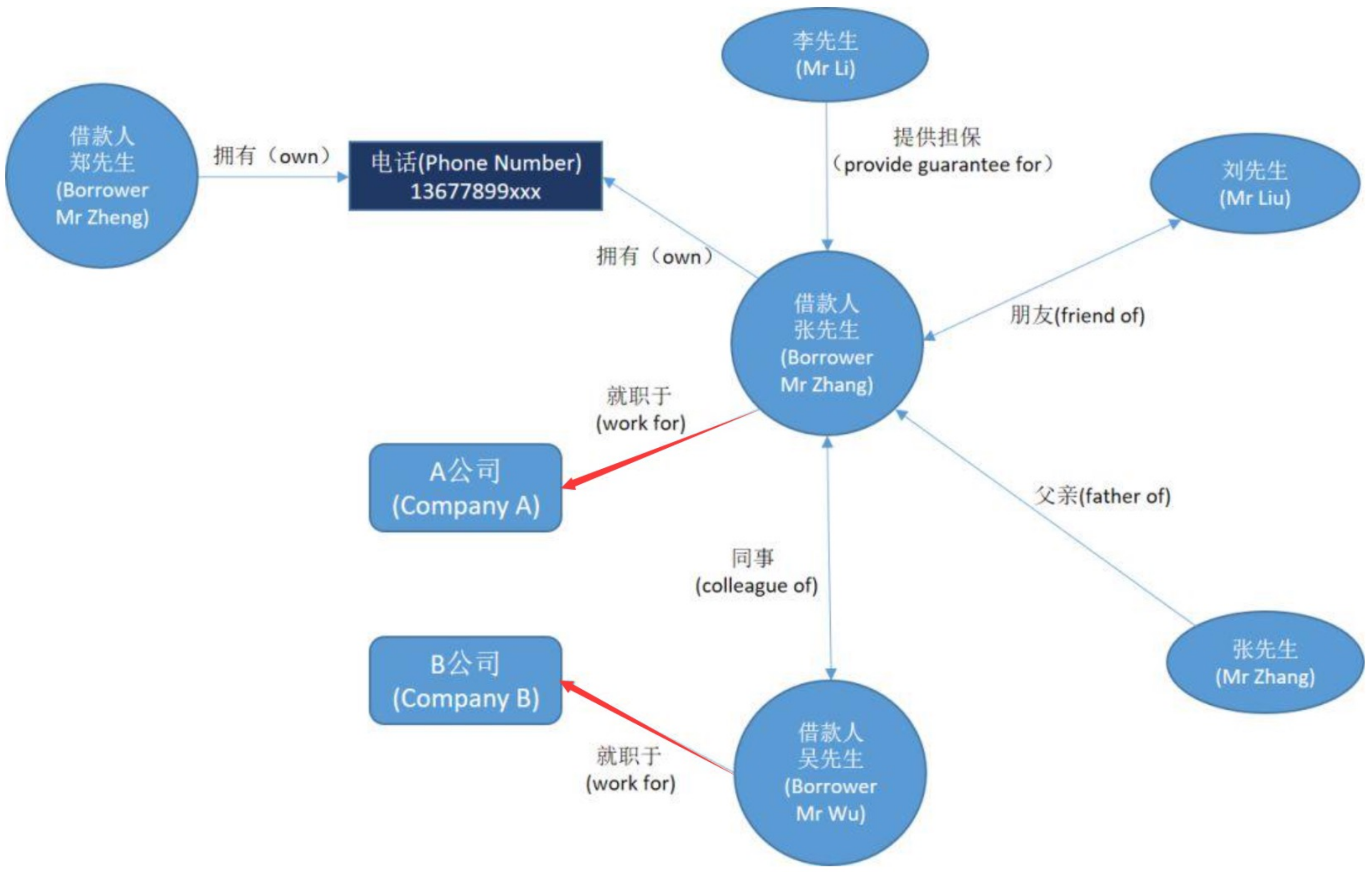
访客A询问机器人“张三的公司在哪里”，之后机器人会根据张三的知识图谱关系，查询到公司在“上海”。



知识图谱应用

举例说明 — 反欺诈情报分析

通过融合来自不同数据源的信息构成知识图谱，同时引入领域专家建立业务专家规则，通过数据不一致性检测，利用绘制出的知识图谱可以识别潜在的欺诈风险。



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

经典算法

Classical

汇报人：杨笔奇、朱逸铭

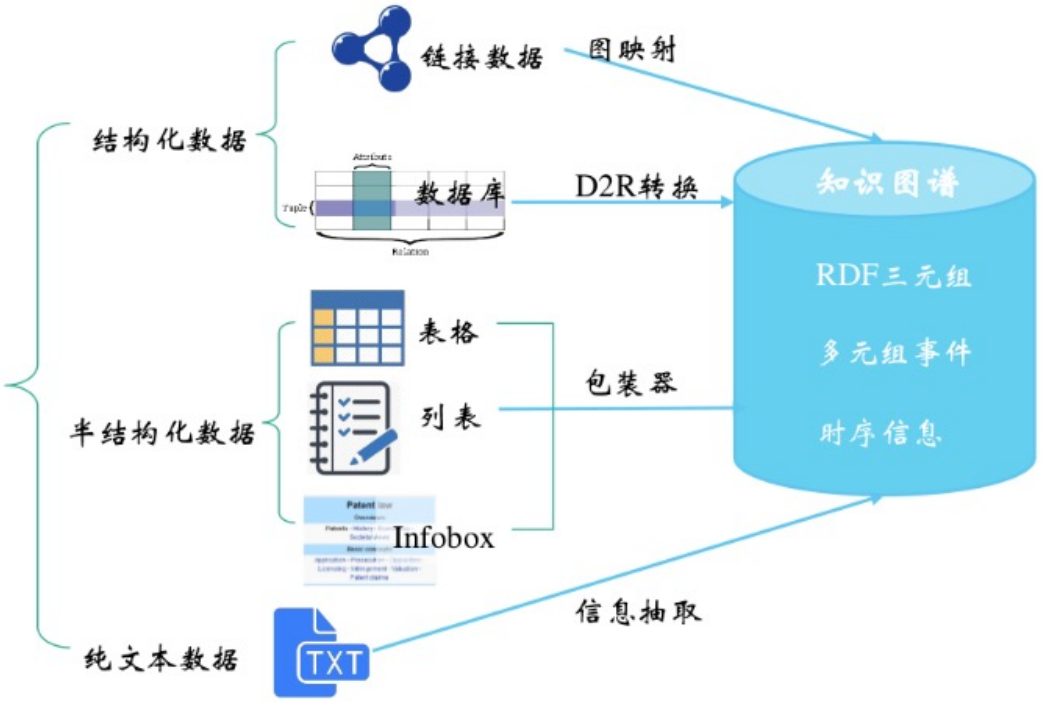
01

知识抽取

Knowledge Extraction

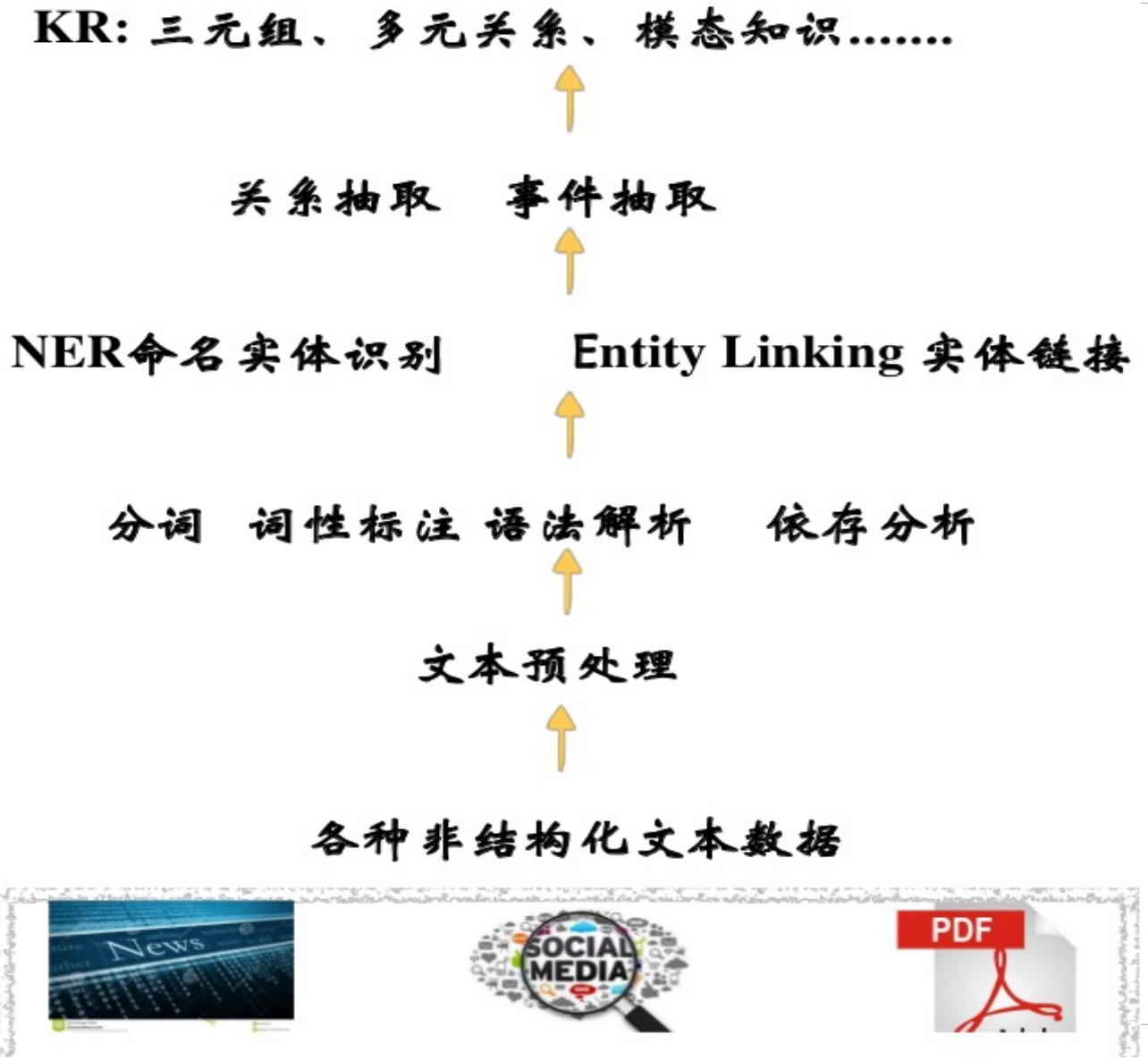
知识抽取

- 知识蕴含在半结构化、非结构化的信息中，需要从中提取出实体、关系、属性等知识要素。
- 涉及的关键技术包括：实体抽取、关系抽取和事件抽取 。



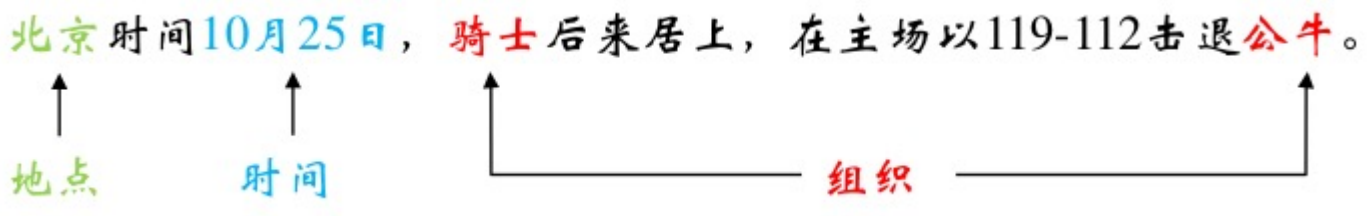
知识抽取

- 知识抽取流程：


KR: 三元组、多元关系、模态知识.....
关系抽取
事件抽取
NER命名实体识别
Entity Linking 实体链接
分词
词性标注
语法解析
依存分析
文本预处理
各种非结构化文本数据
News
SOCIAL MEDIA
PDF

实体抽取

- 实体抽取抽取文本中的原子信息元素，通常包含人名、组织/机构名、地理位置、时间/日期、字符值等标签，具体的标签定义可根据任务不同而调整。如：



- 单纯的实体抽取可作为一个序列标注问题，因此可以使用机器学习中的HMM、CRF、神经网络等方法解决。

关系抽取

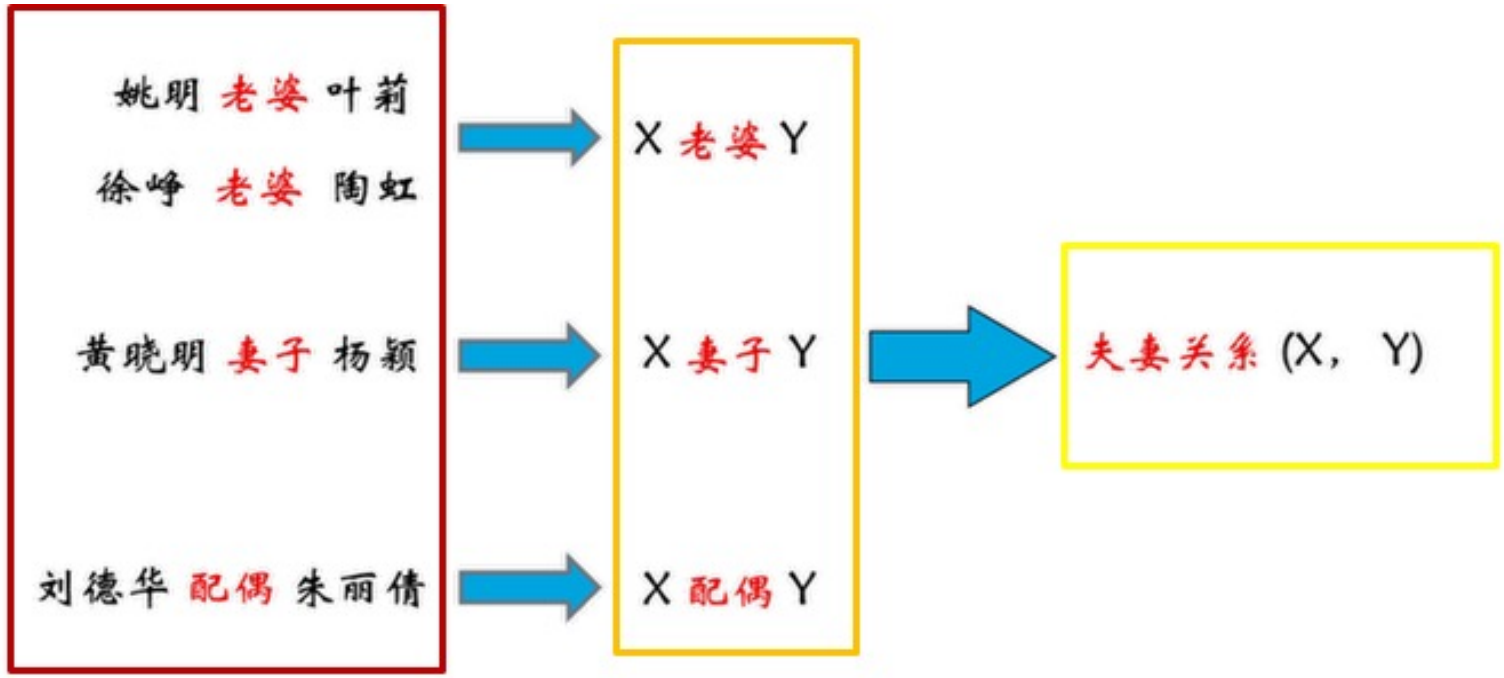
- 关系抽取是从文本中抽取出两个或多个实体之间的语义关系。它是信息抽取研究领域的任务之一。如:

  - 王健林谈儿子王思聪:我期望他稳重一点。
  - 父子(王健林, 王思聪)

- 根据关系抽取方法的不同，可以将其分为:基于模板的方法(触发词的Pattern，依存句法分析的Pattern)、基于监督学习的方法(机器学习方法)、弱监督学习的方法(远程监督、Bootstrapping)。

关系抽取-基于模板的方法

- 早期的关系抽取方法大多基于模板匹配实现。这类方法是基于语言学知识，由领域专家手工编写模板，从文本中匹配具有特定关系的实体。
- 基于触发词的模板：



关系抽取-基于监督学习的方法

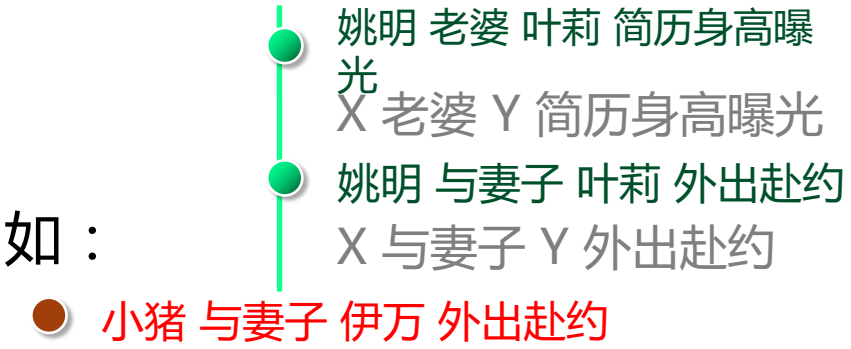
在给定实体对的情况下，根据句子上下文对实体关系进行预测，执行流程为：

- 预先定义好关系的类别。
- 人工标注一些数据。
- 设计特征表示。
- 选择一个分类方法。(SVM、NN、朴素贝叶斯)
- 评估方法。

- 其优点为准确率高，标注的数据越多越准确。缺点为标注数据的成本太高，不能扩展新的关系。
- 近年来：有多个基于深度学习的关系抽取模型被研究者们提出。深度学习的方法不需要人工构建各种特征，其输入一般只包括句子中的词及其位置的向量表示。我们现在已有的基于深度学习的关系抽取方法主要包括流水线方法和联合抽取方法两大类。

关系抽取-基于弱监督学习的方法

- 弱监督学习分为远程监督学习和Bootstrapping：
  - 远程监督学习方法认为若两个实体如果在知识库中存在某种关系,则包含该两个实体的非结构化句子均能表示出这种关系。
  - Bootstrapping
    - 从文档中抽取出包含种子实体的新闻，如：
    - 将抽取出的Pattern去文档集中匹配：
    - 根据Pattern抽取出的新文档如种子库,迭代多轮直到不符合条件



事件抽取

- 事件抽取从自然语言中抽取出用户感兴趣的事件信息,并以结构化的形式呈现出来,例如事件发生的时间、地点、发生原因、参与者等。事件抽取也分为流水线方法和联合抽取方法。

Argument role

事件类型	发布会
公司	苹果公司
时间	西部时间9月12日上午10点
地点	史蒂夫·乔布斯剧院
产品	iPhone8、iPhone7s、iPhone7s plus、apple watch 3、apple TV

arguments

02

知识表示

Knowledge Extraction

知识表示

知识表示研究怎么利用计算机符号来表示人脑中的知识，以及怎么通过符号之间的运算来模拟人脑的推理过程。

早期的知识表示方法：包括一阶谓词逻辑、产生式系统

基于语义网的知识表示框架RDF、RDFS、OWL等

近年来代表模型：距离模型、单层神经网络模型等。

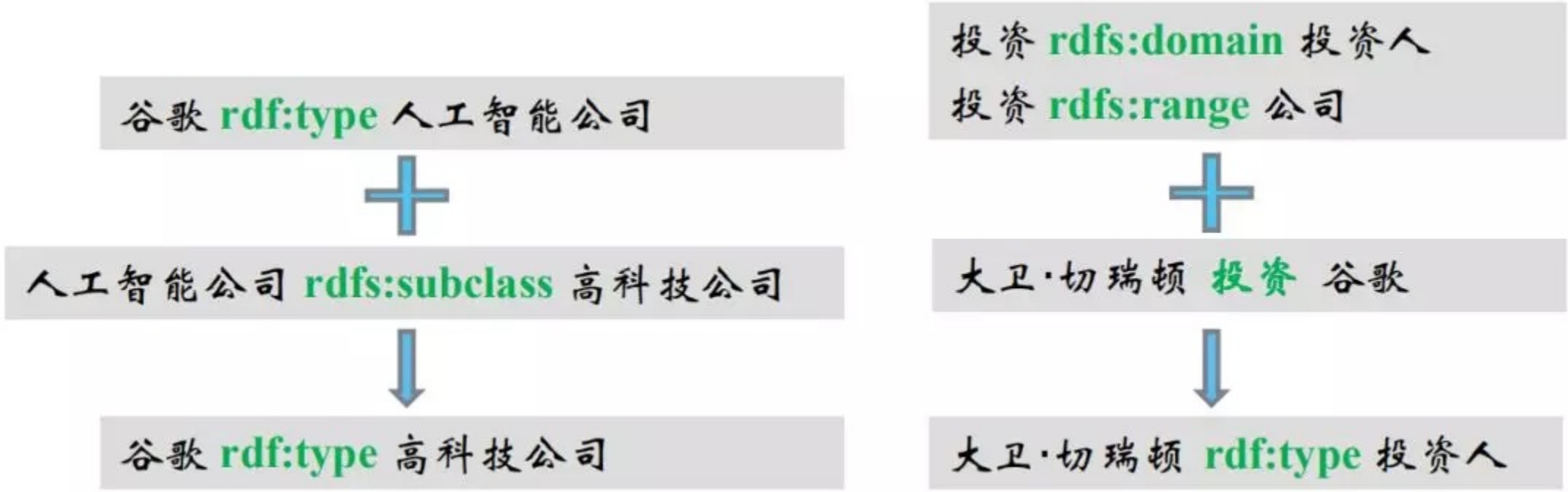
基于语义网的知识表示框架

- RDF：资源描述框架（Resource Description Framework,RDF），一种用于描述Web资源的标记语言，一般采用三元组表示。
- Resource：页面、图片、视频等任何具有URI标识符的资源；
- Decription：属性、特征和资源之间的关系；
- Framework：模型、语言和这些描述的语法；
- RDF是一个三元组（Triple）模型，即每一份知识可以被分解为如下形式：

（subject（主），predicate（谓），object（宾））

基于语义网的知识表示框架

- **RDFS是RDF框架**，在RDF的基础上提供了一个术语、概念等的定义方式，以及哪些属性可以应用到哪些对象上。换句话说，RDFS为RDF模型提供了一个基本的类型系统。常用的Schema词汇有：Class，subClassOf，type，Property，subPropertyOf，Domain，Range。



基于语义网的知识表示框架

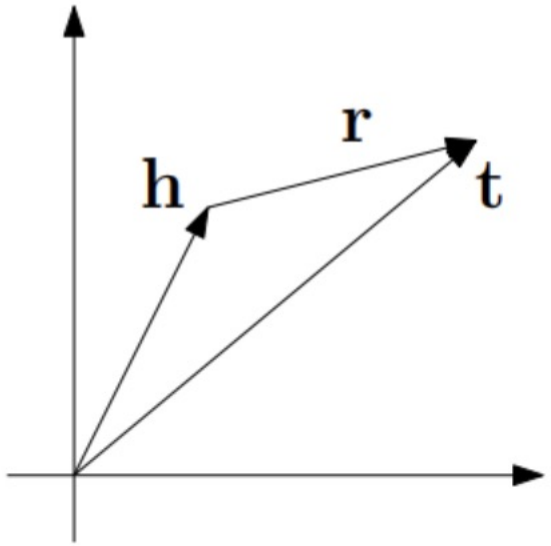
- RDFS在**基数约束、属性特性描述**等方面表达不完整。
- OWL本体语言（OWL Web Ontology Language）去拓展RDF(S)，作为在语义网上表示本体的推荐语言，其目的是为了更好的开发语义网。

```
exp:Person owl:allValuesFrom exp:Women
exp:Person owl:onProperty exp:hasMother
```

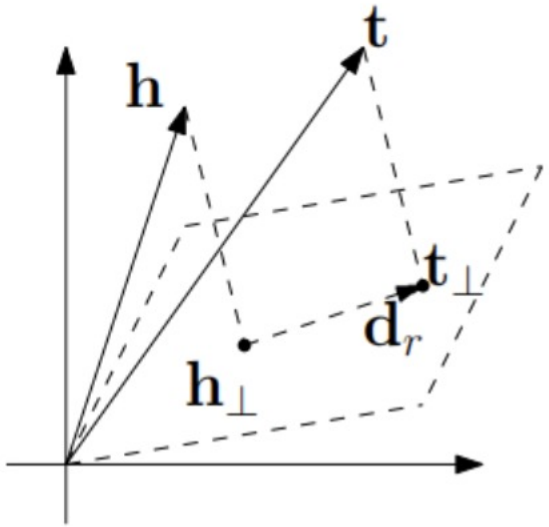
```
exp:SemanticWebPaper owl:someValuesFrom exp:AAAI
exp:SemanticWebPaper owl:onProperty exp:publishedIn
```

知识表示学习——分布式向量

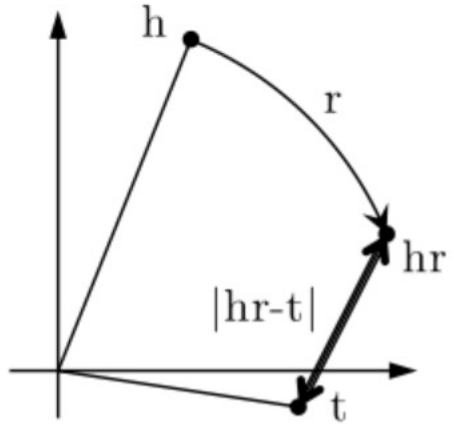
- 表示学习的目标：通过机器学习，将研究对象的语义信息表示为稠密低维实值向量。将实体e和关系r表示为两个不同向量，在向量空间中，通过欧式距离或余弦距离等方式，计算任意两个对象之间的语义相似度。代表模型：距离模型、单层神经网络模型等。



TransE



TransH



RotatE

03

知识融合

Knowledge Fusion

知识融合

知识图谱的构建数据来源十分广泛，不同数据源之间的知识缺乏深入的关联，知识重复问题很严重。知识融合就是将来自不同数据源的异构化、多样化的知识在同一个框架进行消歧、加工、整合等，达到数据、信息以及人的思想等多个角度的融合。

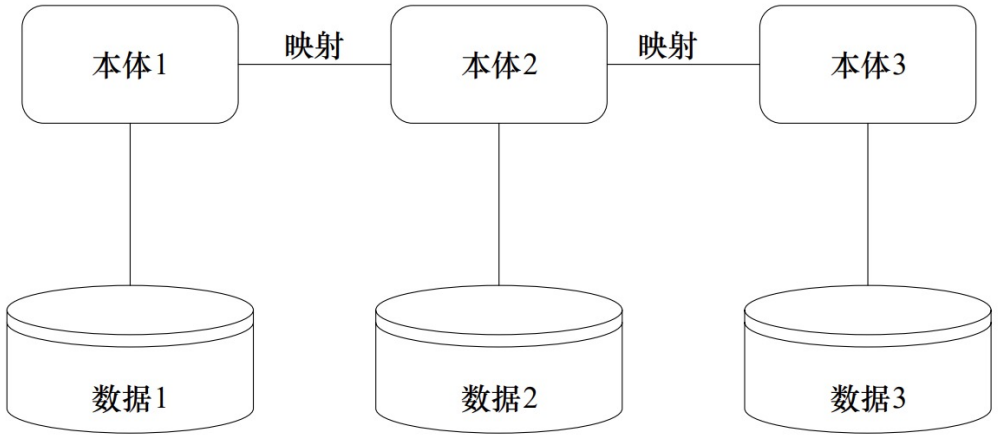
知识融合

当多个知识图谱进行融合 或 将外部关系数据库合并到本体知识库时，需要处理两个层面的问题：

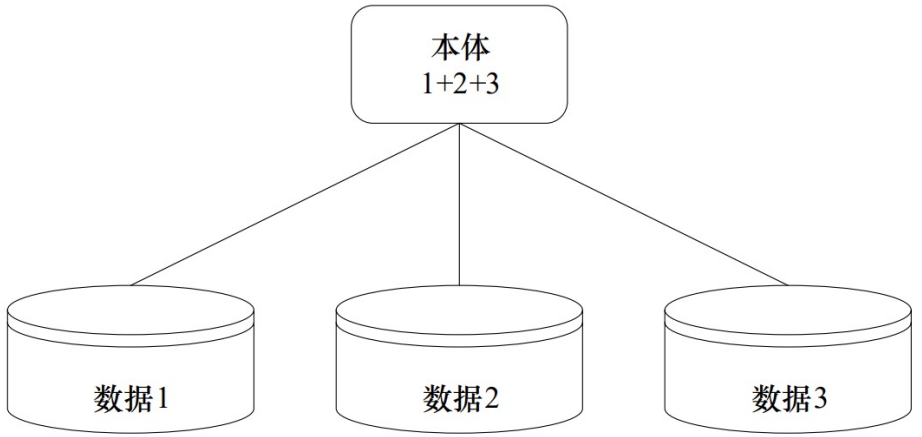
1. 模式层的融合：将新得到的本体融入已有的本体库中，以及新旧本体的融合。

2. 数据层的融合：包括实体的指称、属性、关系以及所属类别等，主要的问题是如何避免实例以及关系的冲突问题，造成不必要的冗余。

模式层的融合



本体映射



本体集成

本体集成是将多个不同数据源的异构本体集成为一个统一的本体；本体映射则是在多个本体之间建立好映射规则使信息在不同的本体之间进行传递。

知识融合

数据层的融合是指实体和关系（包括属性）元组的融合，主要是实体对齐，由于知识库中有些实体含义相同但是具有不同的标识符，因此需要对这些实体进行合并处理。

实体消歧：将形式相同的实体的实际的不同指向区分出来

Bit是一个多义词，请在下列义项上选择浏览（共4个义项） | 收起 ^ 添加义项 +

- **北京理工大学的缩写**
- 百度技术学院简称
- 英文单词
- 中美双边投资协议

实体消歧--词袋模型

MJ：Michael Jordan is a researcher in machine learning

MJ1	researcher	machine	learning

MJ：Michael Jordan plays basketball in Chicago Bulls

MJ2	plays	basketball	Chicago	Bulls

实体消歧—社会化网络

MJ(BasketBall):Pippen,Buckley,Ewing,Kobe

MJ(Machine Learning):Andrew Y. Ng, Nina Balcan

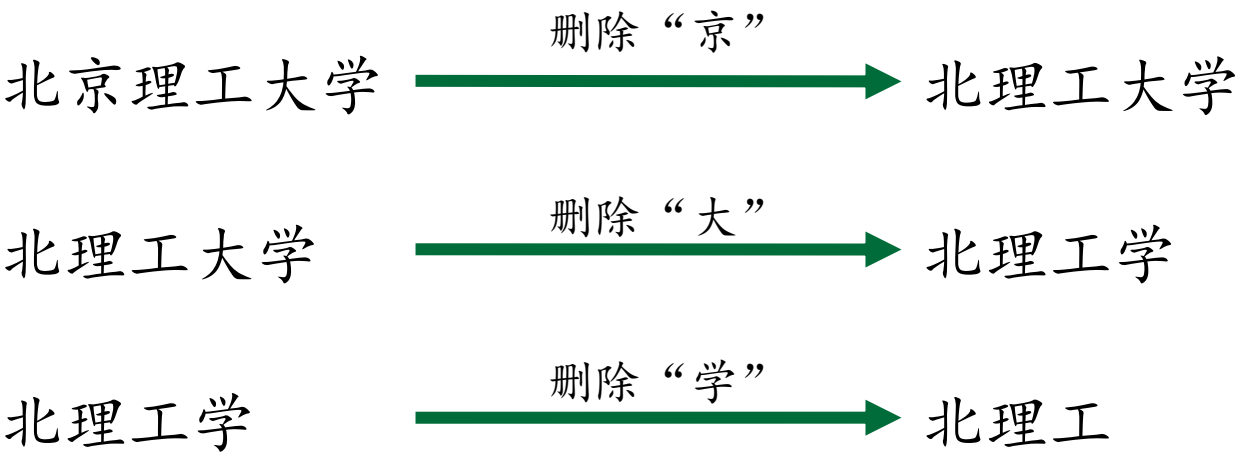
基于社会化网络的实体指称项相似度通常使用基于图的算法，能够充分利用社会化关系的传递性，从而考虑隐藏的关系知识，在某些情况下（特别是结构化数据，eg：论文记录、电影记录等）能够更为准确的实体指称项相似度计算结果。

知识融合 – 实体统一

实体统一：是指判断多个实体是不是属于一个实体

知识融合 – 实体统一

- **Levenshtein Distance（最小编辑距离）**



操作3次！

Levenshtein Distance = 3

- **集合相似度**

$$\mathrm{Dice}(s1,s2)=\frac{2*\mathrm{comm}(s1,s2)}{leng(s1)+leng(s2)}$$

字符串也可以理解为一种集合，因此Dice距离也会用于度量字符串的相似性

基于向量的相似度

TF：词频（term frequency）指的是某一个给定的词语在该文件中出现的频率，衡量了一个词在一个文档中的重要程度

IDF：逆向文件频率（inverse document frequency）是一个词语普遍重要性的度量，如冠词a、an、the等虽然出现频率高但是并不重要

$$TF\text{–}IDF = TF_{i,j} \times IDF_i$$

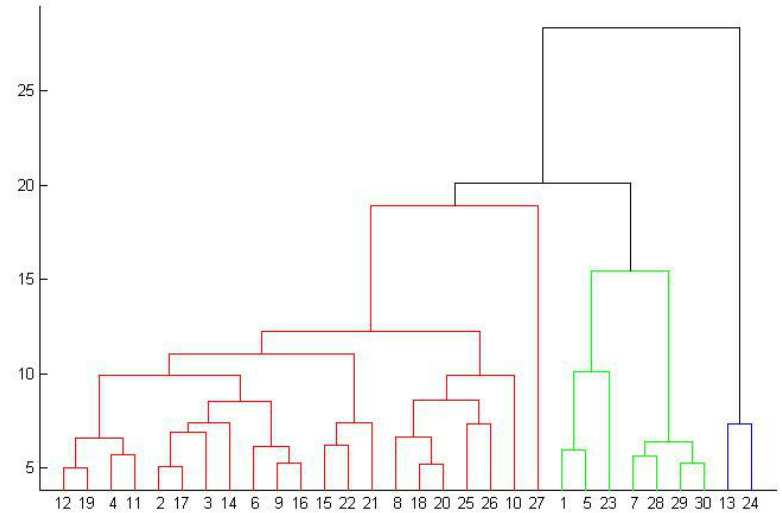
$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$IDF_i = log \frac{|D|}{1 + |j: t_i \in d_j|}$$

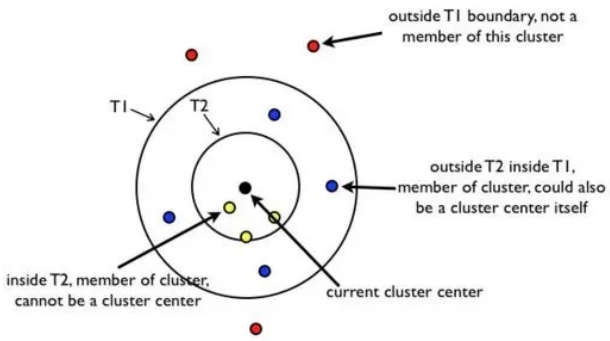
知识融合 –实体统一

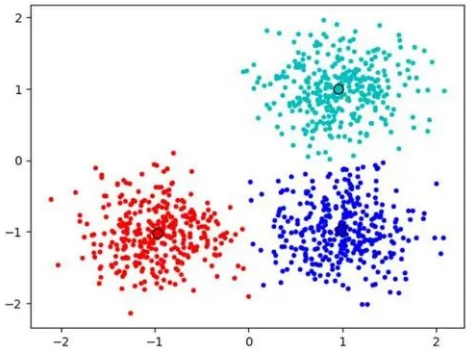
实体对齐：聚类

- 层次聚类



- Canopy + K-Means等





知识融合 – 指代消解

Mention Pair models

将所有的指代词（短语）与所有被指代的词（短语）视作一系列pair，对每个pair二分类决策成立与否。

Mention ranking models

显式地将mention作为query，对所有candidate做rank。

Entity-Mention models

一种更优雅的模型，找出所有的entity及其对话上下文。根据对话上下文聚类，在同一个类中的mention消解为同一个entity。但这种方法其实也用得不多。

04

知识推理

Knowledge Reasoning

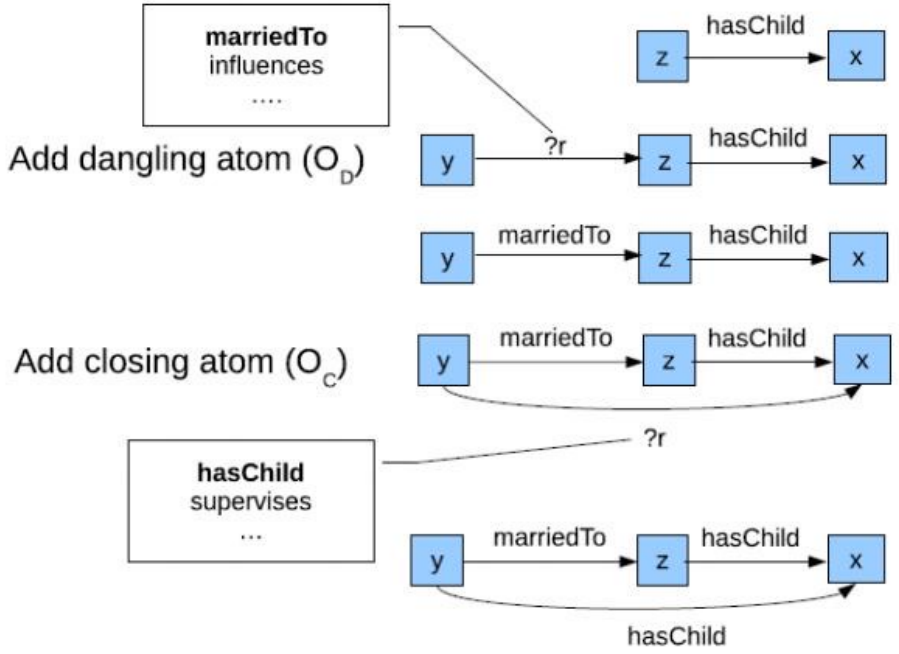
知识推理

知识推理是根据已有的实体关系信息来推断出新的事实结论，进一步丰富知识图谱，满足上游任务的需求。

基于规则的推理

- **添加悬挂边**：悬挂边是指边的一端是一个未出现过的变量，而另一端（变量或常量）是在规则中出现过的
- **添加实例边**：实例边与悬挂边类似，边的一端也是在规则中出现过的变量或常量，但另一端是未出现过的常量，也就是知识库中的实体
- **添加闭合边** ：闭合边则是连接两个已经存在于规则中的元素（变量或常量）的边。

AMIE工作流程示意



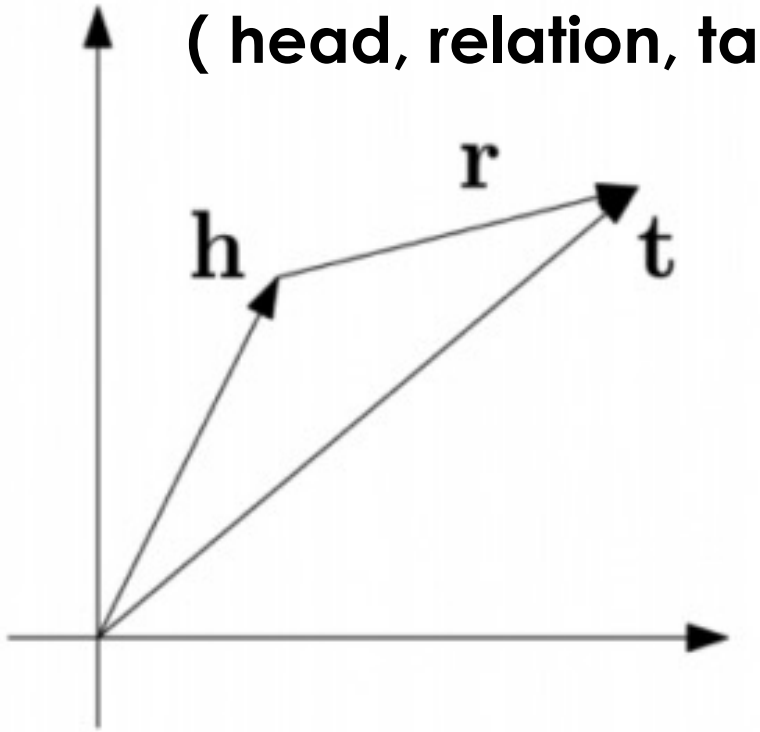
$$\text{marriedTo}(y,z) \wedge \text{hasChild}(y,x) \Rightarrow \text{hasChild}(z,x)$$

基于分布式表示学习的推理

$$\mathcal{L} = \sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in S'_{(h,\ell,t)}} \left[\gamma + d(\boldsymbol{h}+\boldsymbol{\ell},\boldsymbol{t}) - d(\boldsymbol{h'}+\boldsymbol{\ell},\boldsymbol{t'})\right]_+$$

$$S'_{(h,\ell,t)} = \left\{(h',\ell,t)|h' \in E\right\} \cup \left\{(h,\ell,t')|t' \in E\right\}$$

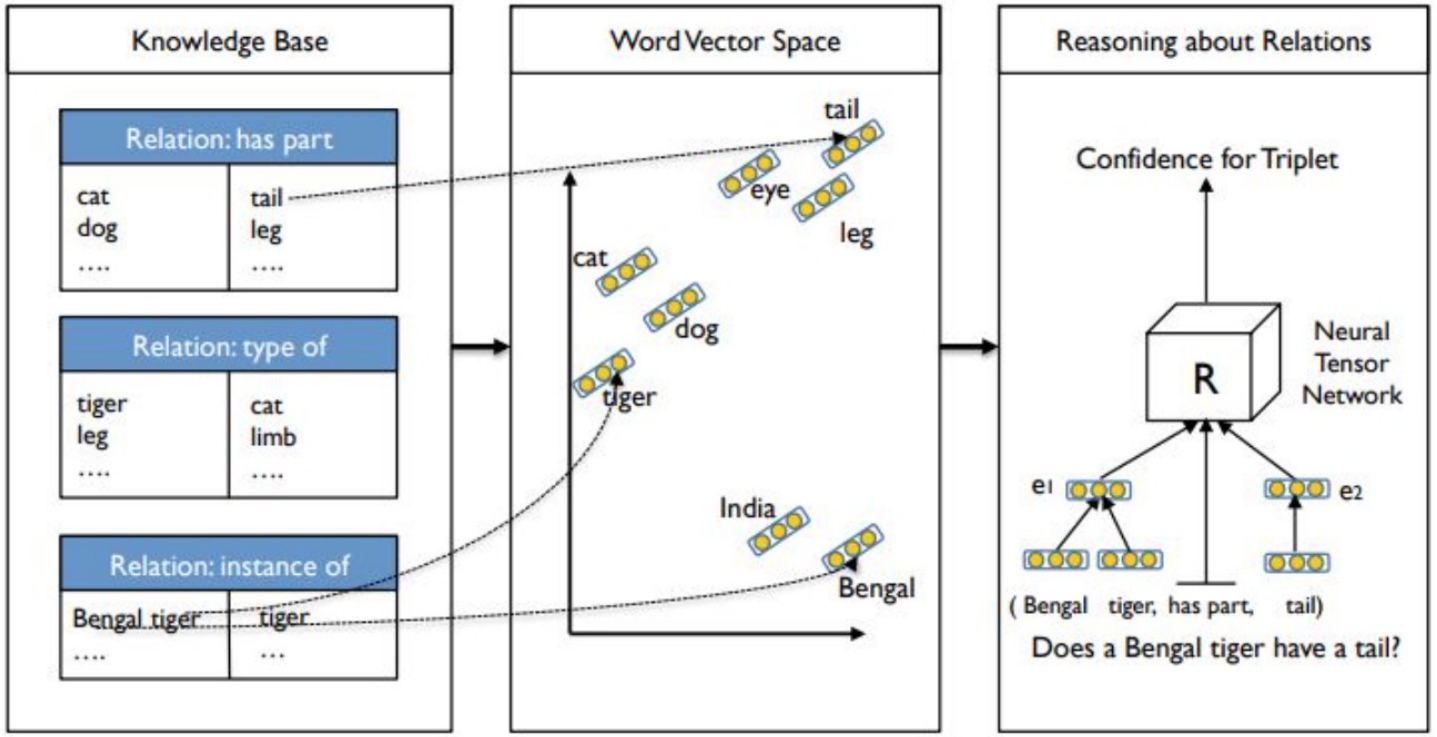
(head, relation, tail)



基于神经网络的推理

在知识推理中，任务是确定两个实体对之间的关系。

对于实体对，选择匹配分数最高的关系。



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

前沿算法

Frontier

汇报人：郭沛祺、张羽冰

前沿进展 – 知识图谱构建中的实体识别

特别领域的数据集，标注的数据量不够？

如医疗领域中，有大量的专业术语需要识别。

“弱监督”

基于规则的弱监督方法：

“Overexpression/amplification of 【靶点】 associated with a worse outcome in patients with 【适应症】 ”

Her-2	NEU; NGL; HER2; TKR1; CD340; HER-2; MLN 19; HER-2/neu; ERRB-2

Breast cancer is the most common malignancy in women in the United States in the year 2000. The proto-oncogene Her-2/ neu (c-erb-B2) has become an increasingly important prognostic and predictive factor in breast cancer. Overexpression/amplification of the Her-2/ neu has been associated with a worse outcome in patients with breast cancer. Herceptin, a “ humanized” murine monoclonal antibody directed against the extracellular domain of the Her-2/ neu protein, is being used to treat breast cancer that overexpresses Her-2/ neu. The status of Her-2/ neu in the tumor has become a critical factor in the management strategy of a breast cancer patient. The objective of this article is to provide a comprehensive review of all aspects of Her-2/ neu in breast cancer, including biology, prognostic and predictive value, targeted Herceptin therapy, and the laboratory testing of Her-2/ neu.

引用： [1] https://journals.lww.com/molecularpathology/Abstract/2001/09000/Her_2__neu_and_Breast_Cancer.1.aspx

前沿进展 – 知识图谱构建中的实体识别（1）

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

《Weakly Supervised Sequence Tagging from Noisy Rules》

论文的主要贡献：

- 提出了Linking Rule的概念

- 提出一个新的生成模型linked hidden Markov models (linked HMMs)

[2] Safranchik E , Luo S , Bach S . Weakly Supervised Sequence Tagging from Noisy Rules[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4):5570-5578.

前沿进展 – 知识图谱构建中的实体识别（1）

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

《Weakly Supervised Sequence Tagging from Noisy Rules》

将规则分为了两类：Tagging Rules & Linking Rules

- Tagging Rules：标注元素序列 —— 给出具体类别
- Linking Rule：判断相邻元素是否是相同标注 —— 只需判断是否同类

In:	She	bikes	the	Washington	Bridge
Out 1:	ABS	ABS	ABS	[I-PER]	ABS
Out 2:	ABS	ABS	ABS	[I-LOC	I-LOC]
Out 3:		ABS	ABS	[ABS	SAME]

[2] Safranchik E , Luo S , Bach S . Weakly Supervised Sequence Tagging from Noisy Rules[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4):5570-5578.

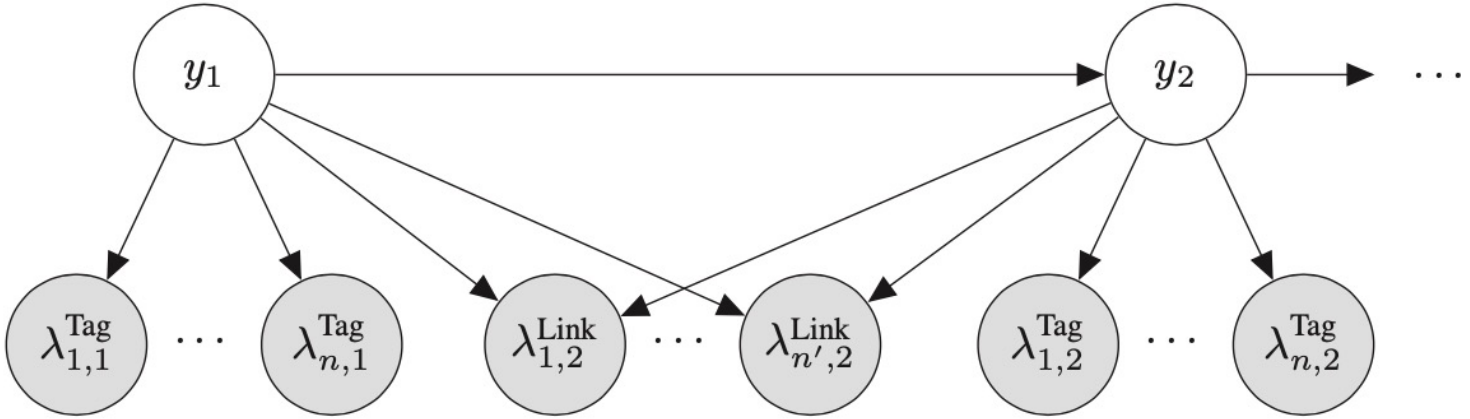
前沿进展 –知识图谱构建中的实体识别（1）

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

《Weakly Supervised Sequence Tagging from Noisy Rules》

Linked HMM是一类动态贝叶斯网络，其主要思想：

- 真实的标签序列表示为潜在随机变量
- 学习一个可以观察 Tagging和Linking Rule 输出的概率生成模型



[2] Safranchik E , Luo S , Bach S . Weakly Supervised Sequence Tagging from Noisy Rules[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4):5570-5578.

前沿进展 – 知识图谱构建中的实体识别（2）

LinkedHMM：186条启发式规则 & 200万个术语词典

“仍然需要大量人工，并对目标数据有深刻的理解。”

[2] Safranchik E , Luo S , Bach S . Weakly Supervised Sequence Tagging from Noisy Rules[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4):5570-5578.

前沿进展 – 知识图谱构建中的实体识别（2）

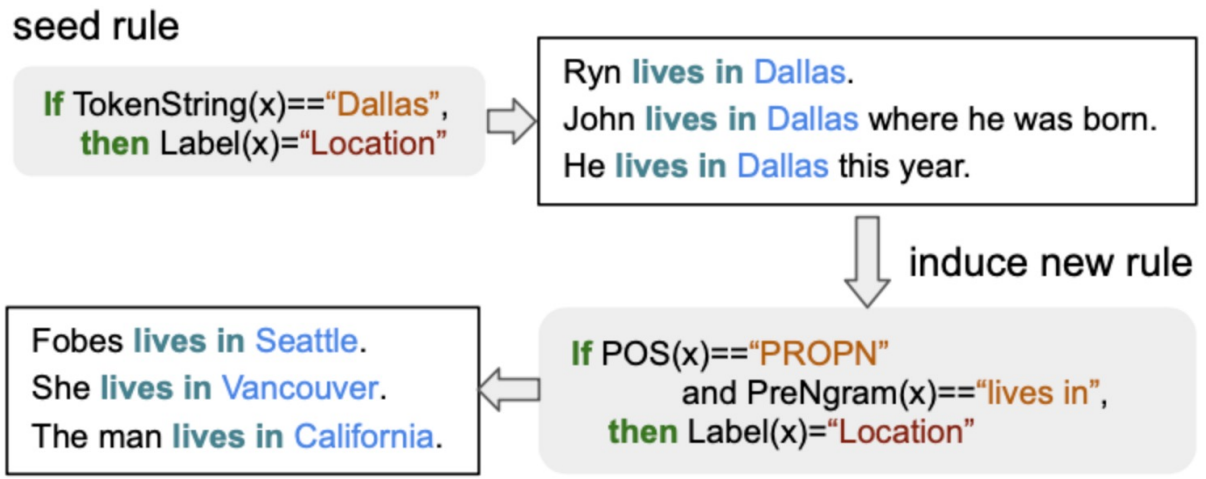
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

《Weakly Supervised Named Entity Tagging with Learnable Logical Rules》

TaLLoR ：

- 一些简单的种子规则（初识设置的简单规则，如20条），从未标记的数据中自动学习其他规则。
- 快速应用到新兴领域或者特别定制的实体类型当中。
- 学习的规则是可解释的，对于非专业人士可以针对错误预测进行修改。

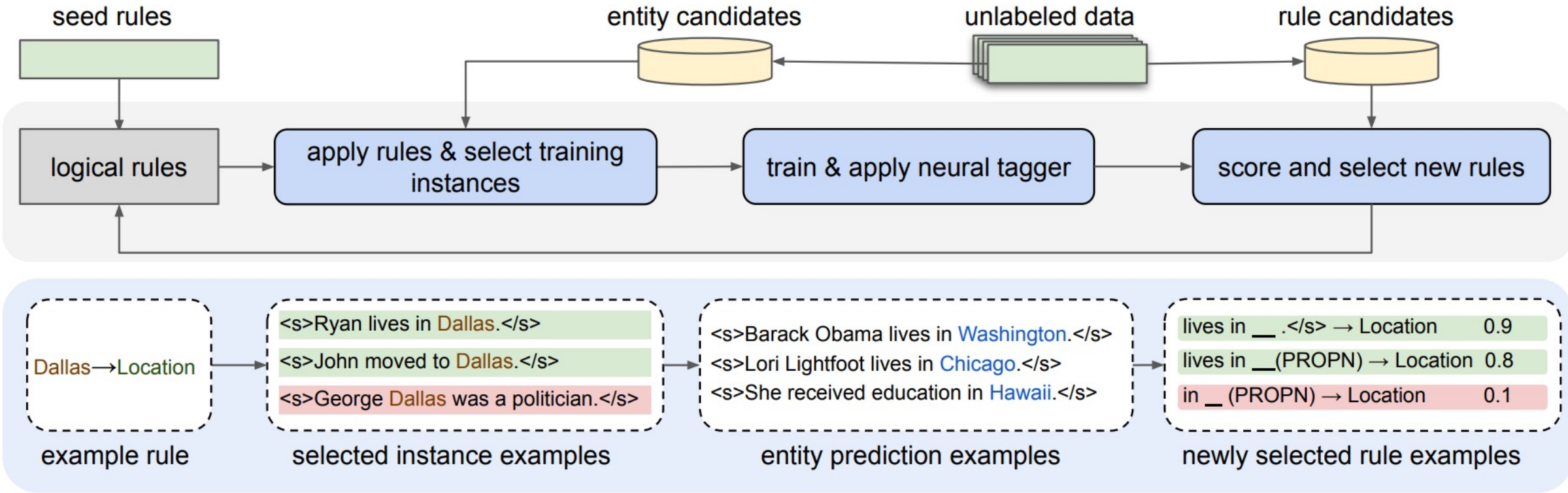
“更适用于工业界”



[3] Jiacheng Li, Haibo Ding, Jingbo Shang, Julian McAuley, Zhe Feng. Weakly Supervised Named Entity Tagging with Learnable Logical Rules. (ACL 2021)

前沿进展 – 实体识别（2）

整体框架和示例



[3] Jiacheng Li, Haibo Ding, Jingbo Shang, Julian McAuley, Zhe Feng. Weakly Supervised Named Entity Tagging with Learnable Logical Rules. (ACL 2021)

前沿进展 – 知识图谱构建中的实体识别（2）

TaLLoR 自动学习的两个关键点：

- 如何根据规则同时探测实体边界和预测实体类型。
- 如何根据规则生成正确且多样的标签。

5种简单规则：

1. TokenString（将字符串分割成类似与人类语言的词块）与词法string相匹配。
2. PreNgram 匹配前文的文本tokens。
3. PostNgram匹配后文的文本tokens。
4. POSTag匹配词性标签。
5. DependencyRel匹配中心词（head word）的依赖关系

“John lives in Dallas where he was born”

↓

“lives in ”匹配多个预测区间

“Dallas”, “Dallas where”, “Dallas where he” etc.

↓

使用前文线索规则+词性检测（名词）

↓

“Dallas”

前沿进展 – 实体识别（2）

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

TaLLoR 的神经标签网络：

给定一个span和对应的句子：

- 使用预训练语言模型编码所有tokens。
- 使用Bi-LSTM和self-attention获取上下文嵌入。
- 计算Span的嵌入。
- 多层感知机预测标签。

逻辑规则打分：

规则正确率

规则覆盖面

$$F(r) = \frac{F_i}{N_i} \log_2(F_i)$$

F_i：用规则r预测标签为i的span个数

N_i：所有用规则r匹配的span个数

[3] Jiacheng Li, Haibo Ding, Jingbo Shang, Julian McAuley, Zhe Feng. Weakly Supervised Named Entity Tagging with Learnable Logical Rules. (ACL 2021)

前沿进展 – 关系抽取

- 文档级关系抽取
- 开放领域关系抽取
- 对于关系抽取，重要的信息是什么？

前沿进展 – 关系抽取（1）

- 文档级关系抽取：Entity-Relation Extraction as Multi-turn Question Answering
  - 思想：将关系抽取转化为机器阅读理解问题，即根据上下文预测答案区间。
  - 模型概述
    - 提取头实体：将每个实体类型转化为一个问题，并通过回答问题来提取头实体
    - 提取关系和尾实体：为每个关系类型设计问题模板并人工定义一个关系链，按照顺序执行多轮问答，得到关系和尾实体

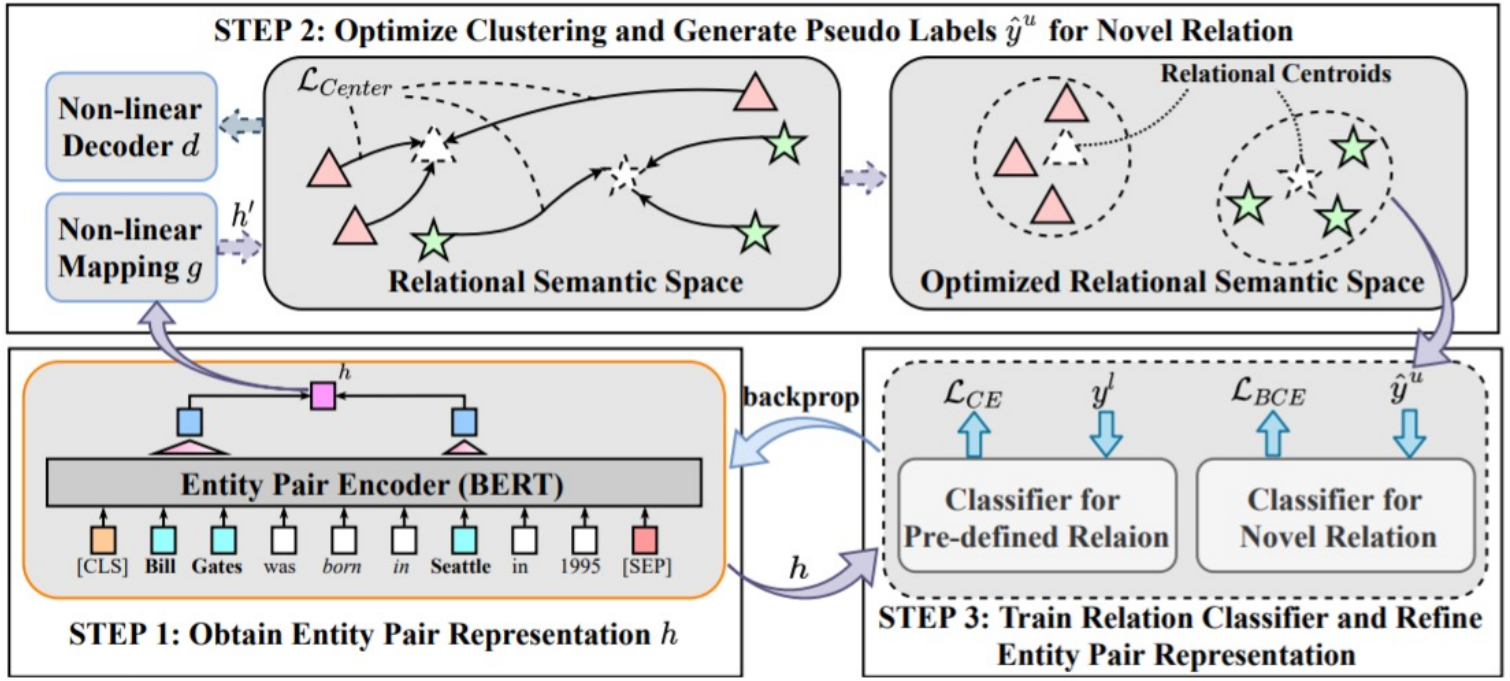
Li X, Yin F, Sun Z, et al. Entity-relation extraction as multi-turn question answering[J]. arXiv preprint arXiv:1905.05529, 2019.

前沿进展 – 关系抽取（2）

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 开放领域关系抽取：A Relation-Oriented Clustering Method for Open Relation Extraction

思想：给定一个预定义关系的数据集(IND)和一个未知关系的数据集(OOD)，用一个面向关系的聚类模型来识别未标记数据中的新关系



[6] Zhao J, Gui T, Zhang Q, et al. A Relation-Oriented Clustering Method for Open Relation Extraction[J]. arXiv preprint arXiv:2109.07205, 2021.

前沿进展 – 关系抽取（3）

- 对于关系抽取，重要的信息是什么？
  - Learning from Context or Names? An Empirical Study on Neural Relation Extraction

- C+M vs C+T：说明实体类型信息比实体提及更重要
- C+T vs OnlyT：说明上下文信息也很重要

Model	C+M	C+T	OnlyC	OnlyM	OnlyT
CNN	0.547	0.591	0.441	0.434	0.295
BERT	0.683	0.686	0.570	**0.466**	0.277
MTB	**0.691**	**0.696**	**0.581**	0.433	**0.304**

[7] Peng H, Gao T, Han X, et al. Learning from context or names? an empirical study on neural relation extraction[J]. arXiv preprint arXiv:2010.01923, 2020.

前沿进展 – PLMs是下一代知识图谱吗

- 从PLMs中提取知识
  - 利用大规模预训练模型可以直接获取事实型知识
  - 不需要人工标注，支持开放域查询

前沿进展 – PLMs是下一代知识图谱吗

Question: Beijing is the capital of [MASK].
(北京是哪里的首都)
BERT: Beijing is the capital of China.
(北京是中国的首都)

Question: Pride and Prejudice is written in [MASK].
(傲慢与偏见是用什么语言写的)
BERT: Pride and Prejudice is written in English.
(傲慢与偏见是用英文写的)

Question: The desk have [MASK] eyes.
(桌子有几只眼睛)
BERT: The desk have no eyes.
(桌子没有眼睛)

Question: The color of the dove is [MASK].
(鸽子是什么颜色的)
BERT: The color of the dove is white.
(鸽子是白色的)

前沿进展 – PLMs是下一代知识图谱吗

- PLMs在有些问题的回答上不尽人意

Question: A man is [MASK] than a house.
(一个人比一栋房子要？)
BERT: A man is bigger than a house.
(一个人比一栋房子要大)

Question: A paper have [MASK] legs.
(一张纸有几条腿？)
BERT: A paper have two legs.
(一张纸有两条腿)

Question: The [MASK] have fur and claws.
(什么动物有毛和爪子)
BERT: The males have fur and claws.
(男性拥有毛和爪子)

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

5 Demo

郭沛祺 张羽冰

Demo(1) - 结构化知识构建图谱

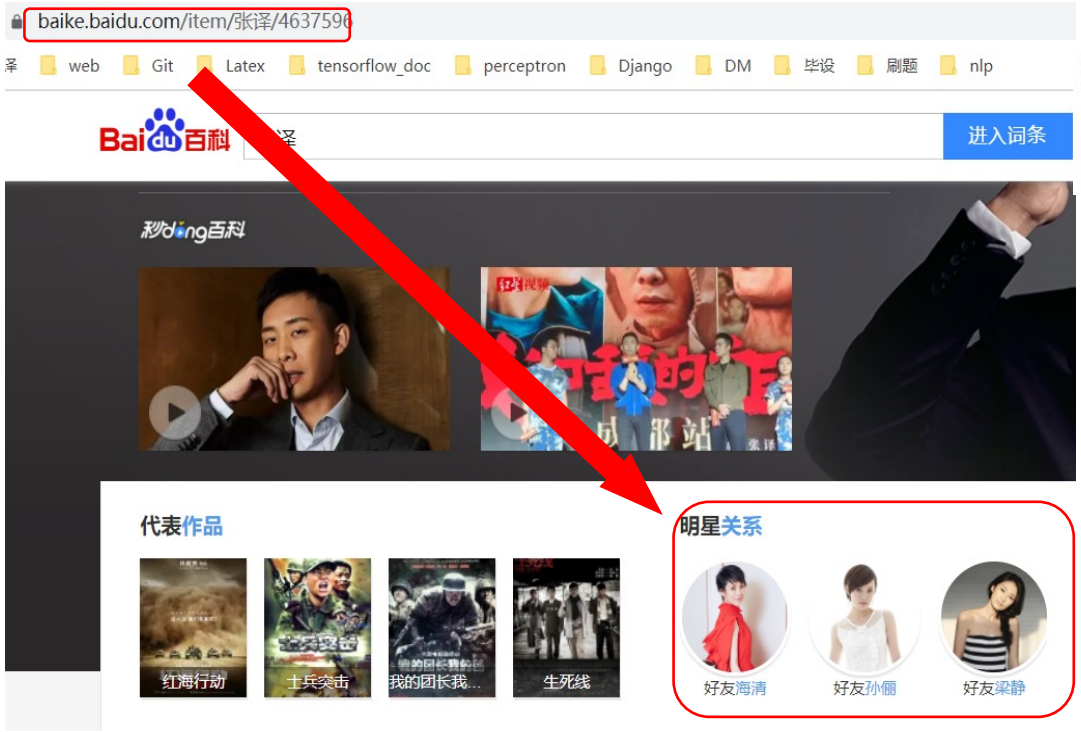
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 影视明星关系知识图谱构建
  - 使用python的urllib库编写爬虫程序，使用xpath解析网页，从互联网爬取数据，得到百度百科演员详情页的url

Demo(1) - 结构化知识构建图谱

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 影视明星关系知识图谱构建
  - 根据演员详情URL从百度百科爬取演员人物关系，返回人物关系三元组



张译

['好友', '好友', '好友', '好友', '好友', '好友', '好友', '好

['海清', '孙俪', '梁静', '陈建斌', '郝蕾', '兰晓龙', '王宝强'

```
33  张译,八佰,参演,1
34  张译,梁静,好友,1
35  张译,海清,好友,1
36  张译,陈建斌,好友,1
37  张译,郝蕾,好友,1
38  张译,兰晓龙,好友,1
39  张译,王宝强,好友,1
40  张译,廖凡,好友,1
41  张译,黄渤,好友,1
42  张译,王珞丹,好友,1
43  张译,李晨,好友,1
44  张译,孙俪,好友,1
```

Demo(1) - 结构化知识构建图谱

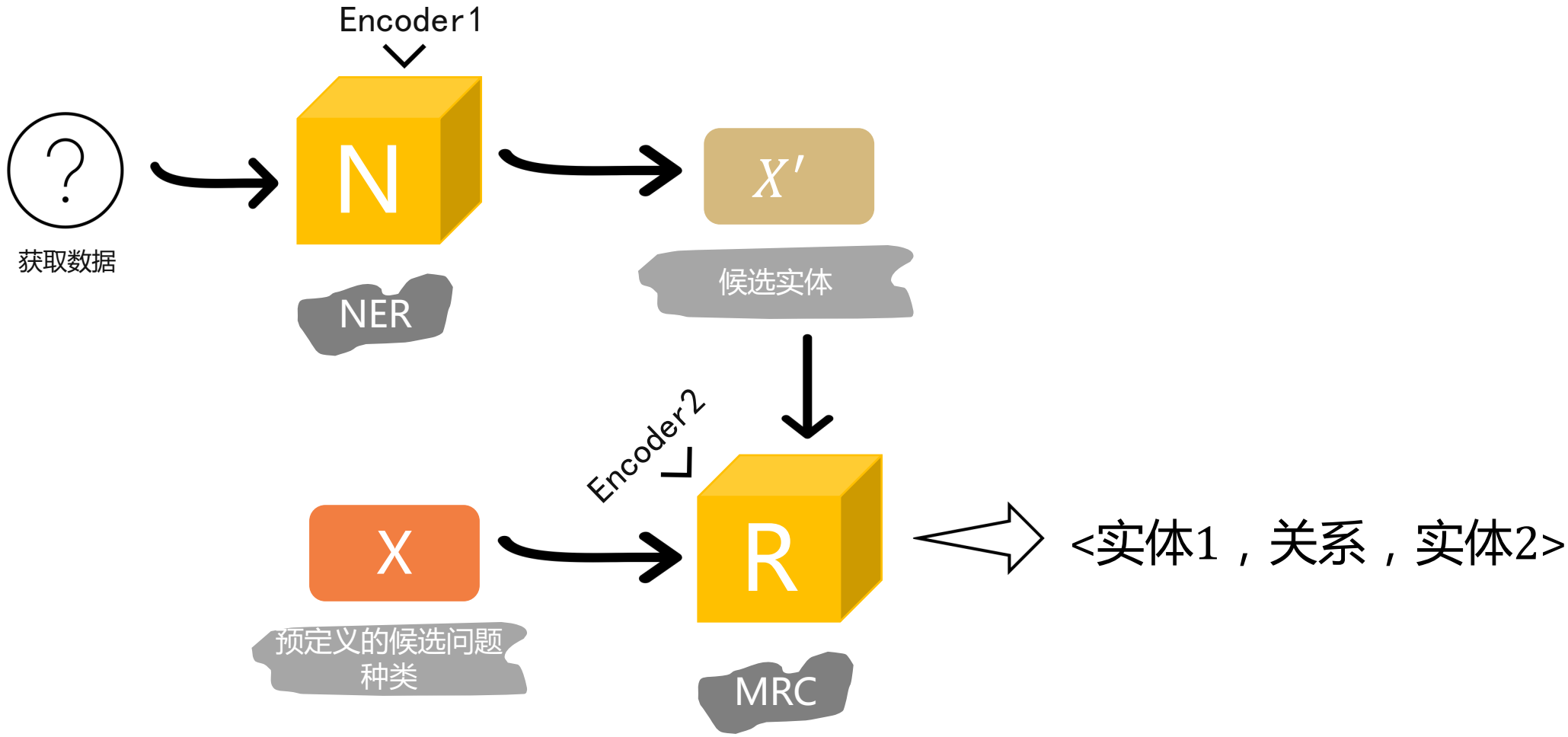
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 影视明星关系知识图谱构建

DEMO(2) – 非结构化数据构建知识图谱

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

模型整体框架流程

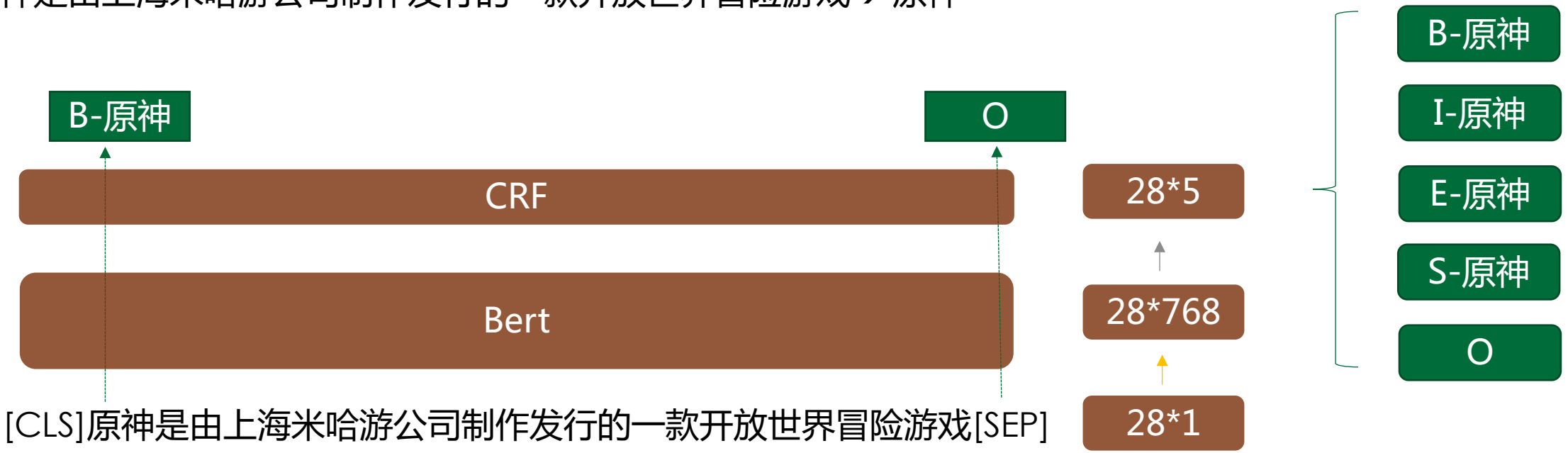


DEMO(2) – 非结构化数据构建知识图谱

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 使用命名实体识别模型检测潜在的实体
  - 使用序列标注模型

原神是由上海米哈游公司制作发行的一款开放世界冒险游戏→ 原神



DEMO(2) – 非结构化数据构建知识图谱

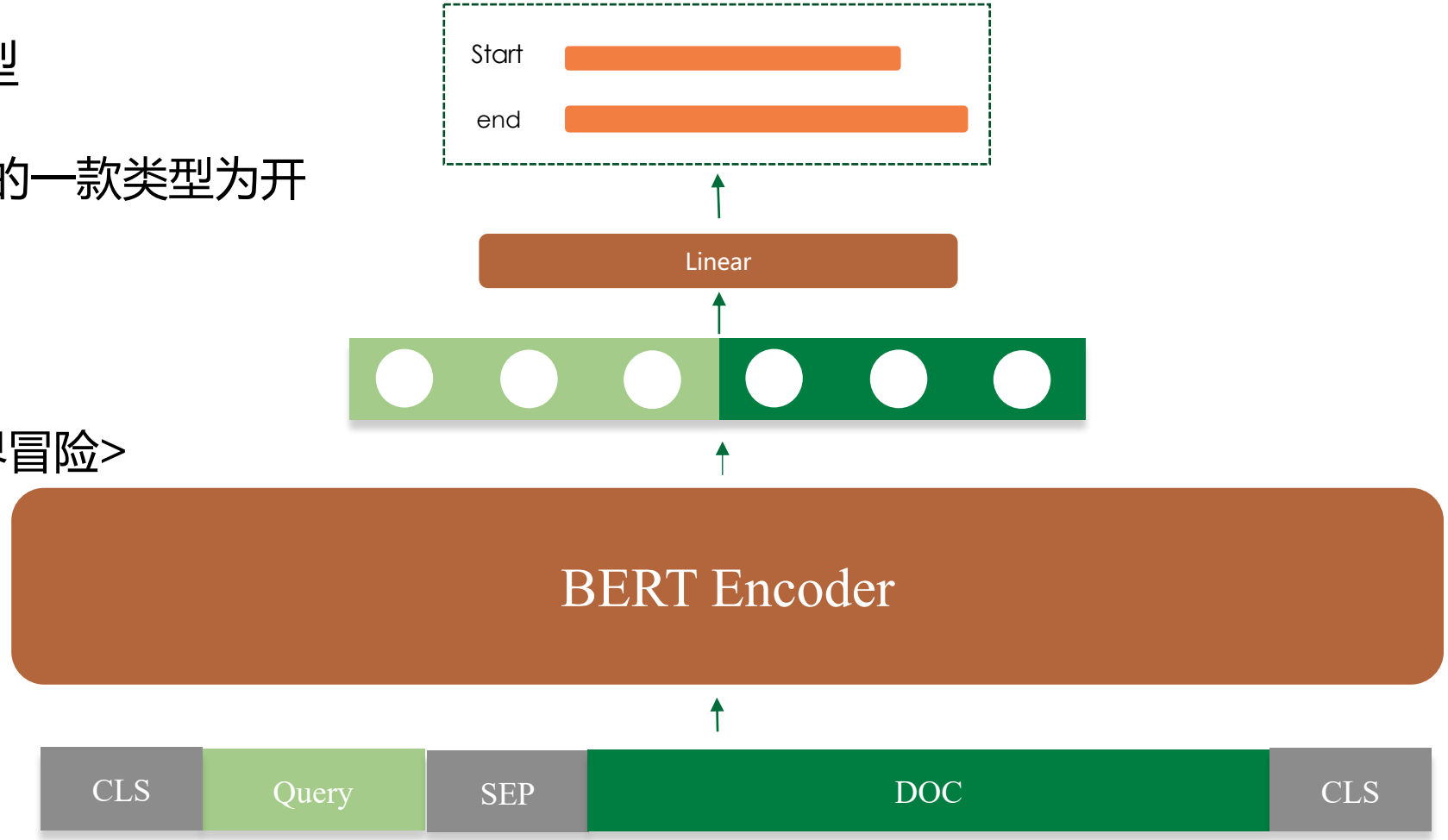
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 使用【关系抽取模型】的抽取潜在关系
  - 使用机器阅读理解模型

原神是由上海米哈游制作发行的一款类型为开放世界冒险的游戏

<原神，制作，上海米哈游>

<原神，游戏类型，开放世界冒险>



DEMO(2) – 非结构化数据构建知识图谱

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

《原神》是由上海米哈游制作发行的一款类型为开放世界冒险的游戏，于2017年1月底立项，原初测试于2019年6月21日开启，再临测试于2020年3月19日开启，启程测试于2020年6月11日开启，PC版技术性开放测试于9月15日开启，公测于2020年9月28日开启。在数据方面，同在官方服务器的情况下，iOS、PC、Android平台之间的账号数据互通，玩家可以在同一账号下切换设备。

原神是哪个公司制作的？

原神是什么类型的游戏？

原神什么时间立项？

原神什么时间原初测试？

原神在什么平台？

DEMO(2) – 非结构化数据构建知识图谱

游戏发生在一个被称作“提瓦特”的幻想世界，在这里，被神选中的人将被授予“神之眼”，导引元素之力。玩家将扮演一位名为“旅行者”的神秘角色，在自由的旅行中邂逅性格各异、能力独特的同伴们，和他们一起击败强敌，找回失散的亲人——同时，逐步发掘“原神”的真相。

原神发生在什么世界？

原神中，被选中的人会被授予什么？

原神中，玩家将扮演什么角色？

DEMO(2) – 非结构化数据构建知识图谱

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 原神是哪个公司制作的？
- <原神，制作，上海米哈游>


```
"1": [
    {
        "text": "上海米哈游",
        "probability": 0.9602254368524571,
        "start_logit": 7.576770782470703,
        "end_logit": 7.91173791885376
    },
    {
        "text": "上海米哈游制作发行",
        "probability": 0.023188708346596098,
        "start_logit": 7.576770782470703,
        "end_logit": 4.188235282897949
    },
    {
        "text": "由上海米哈游",
        "probability": 0.006813419952539035,
        "start_logit": 2.6284968852996826,
        "end_logit": 7.91173791885376
    },
    {
        "text": "上海米哈游制作发行的一款类型为开放世界冒险的游戏",
        "probability": 0.0053622870243267935,
        "start_logit": 7.576770782470703,
        "end_logit": 2.7239603996276855
    },
    {
        "text": "上海米哈游制作发行的",
        "probability": 0.0016410091656459665,
        "start_logit": 7.576770782470703,
        "end_logit": 1.5398812294006348
    },
    {
```

DEMO(2) –非结构化数据构建知识图谱

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

各个问题的答案

原神是哪个公司制作的？


```
"text": "上海米哈游",
"probability": 0.9602254368524571,
```


原神是什么类型的游戏？


```
"text": "开放世界冒险的游戏",
"probability": 0.9855354076889581,
```


原神什么时间立项？


```
"text": "2017年1月底",
"probability": 0.9012254418101031,
```


原神什么时间原初测试？


```
"text": "2019年6月21日开启",
"probability": 0.66459284207094,
```


原神在什么平台？


```
"text": "iOS、PC、Android平台",
"probability": 0.8538383621046216,
```


原神发生在什么世界？


```
"text": "一个被称作"提瓦特"的幻想世界",
"probability": 0.36526204124876765,
```


原神中，被选中的人会被授予什么？


```
"text": ""神之眼"",
"probability": 0.9615798684390593,
```

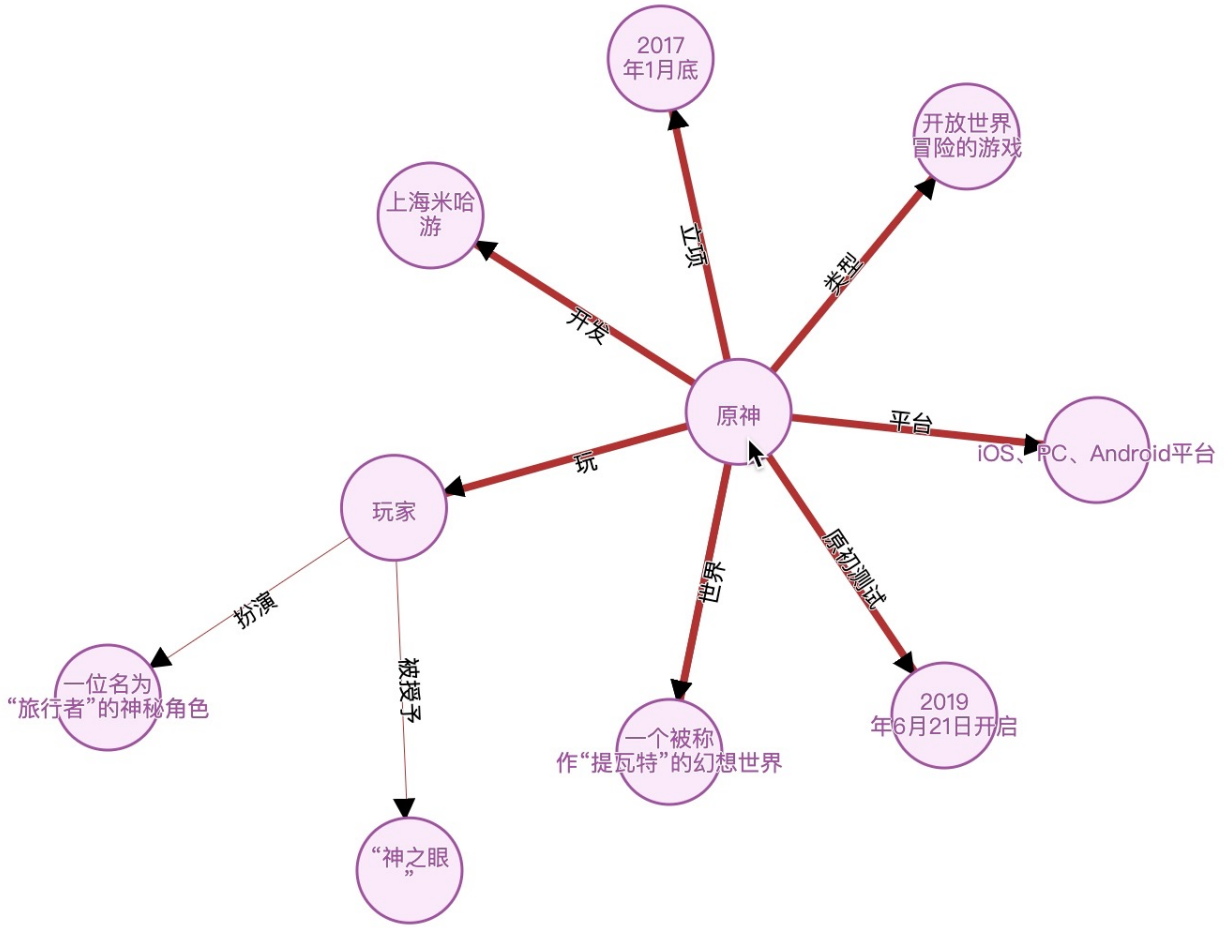

原神中，玩家将扮演什么角色？


```
"text": "一位名为"旅行者"的神秘角色",
"probability": 0.26972539073693885,
```

DEMO(2) – 非结构化数据构建知识图谱

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 构建可视化知识图谱



谢谢大家
敬请批评指正

Thanks for your listening

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

时间：2021.11.15