本章主要内容

1. 知识与知识表示的概念
2. 一阶谓词逻辑表示法
3. 产生式表示法
4. 框架表示法
5. 知识图谱

产生式

- “产生式”由美国数学家波斯特（E.POST）在1934年首先提出，它根据串代替规则提出了一种称为波斯特机的计算模型，模型中的每条规则称为产生式。
- 许多专家系统用它来表示知识
- 产生式通常用于表示事实、规则以及它们的不确定性度量，适合于表示事实性知识和规则性知识。

产生式

- （1）确定性规则知识的产生式表示

- 基本形式： IF P THEN Q

  或者： $P \rightarrow Q$

  P 是产生式的前提，Q 是一组结论或操作。

  如果前提P被满足，则可得到结论Q或执行Q所规定的操作。

例如：

r_1：IF 动物会飞 AND 会下蛋 THEN 该动物是鸟

编号　　*P*　　*Q*

产生式

- （2）不确定性规则知识的产生式表示

- 基本形式： IF *P* THEN *Q*（置信度）

  或者：*P* →*Q*（置信度）

例如：IF 发烧 THEN 感冒（0.6）

例如在专家系统MYCIN中：

IF 本微生物的染色斑是革兰氏阴性

本微生物的形状呈杆状

病人是中间宿主

Then 该微生物是绿脓杆菌，置信度是0.6

产生式

- （3）确定性事实性知识的产生式表示

  - 三元组表示：（对象，属性，值）

    或者：（关系，对象1，对象2）

    例： 老李年龄是40岁： （*Li*，*age*，40）

    老李和老王是朋友：（*friend*，*Li*，*Wang*）

产生式

- （4）不确定性事实性知识的产生式表示

- 四元组表示：（对象，属性，值，置信度）

  或者：（关系，对象1，对象2，置信度）

例： 老李年龄很可能是40岁： （*Li*，*age*，40，0.8）

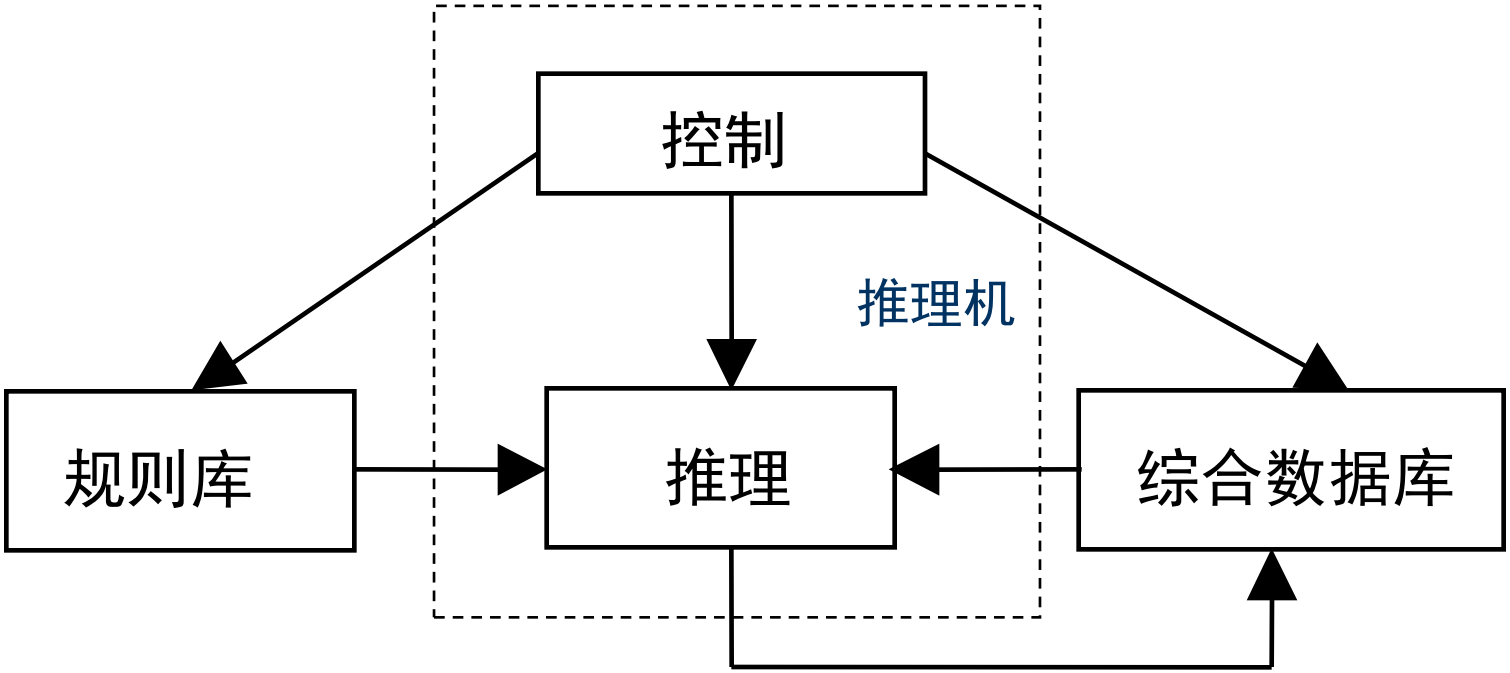
老李和老王不大可能是朋友：（*friend*，*Li*，*Wang*，*0.1*）

产生式和蕴含式的区别

- 产生式与谓词逻辑中的蕴含式基本形式相同
- 蕴含式是产生式的一种特殊情况
  - 除逻辑蕴含外，产生式还包括各种操作、规则、变换、算子、函数等。例如，“如果炉温超过上限，则立即关闭风门”是一个产生式，但不是蕴含式。产生式的外延很广。
  - 蕴含式只能表示精确知识，而产生式不仅可以表示精确的知识，还可以表示不精确知识。蕴含式的匹配总要求是精确的。产生式匹配可以是精确的，也可以是不精确的，只要按某种算法求出的相似度落在预先指定的范围内就认为是可匹配的。
- 产生式的形式描述及语义：
  - 巴科斯范式BNF（Backus normal form）

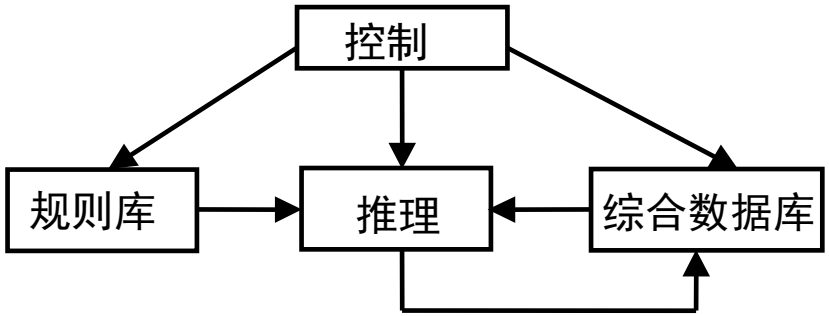
产生式系统

- 一组产生式放在一起，互相配合，一个产生式的结论可以供另一个作为事实使用，以求得问题的解。



产生式系统

- 规则库：用于描述相应领域内知识的产生式集合。
- 综合数据库（事实库、上下文、黑板等）：一个用于存放问题求解过程中各种当前信息的数据结构。
- 控制系统（推理机）：由一组程序组成，负责整个产生式系统的运行，实现对问题的求解。
  - 推理
  - 冲突消解
  - 执行规则
  - 检查推理终止条件



产生式系统例子

- 例如：动物识别系统——识别虎、金钱豹、斑马、长颈鹿、鸵鸟、企鹅、信天翁七种动物的产生式系统。







产生式系统例子

r_1: IF 该动物有毛发 THEN 该动物是哺乳动物

r_2: IF 该动物有乳房 THEN 该动物是哺乳动物

r_3: IF 该动物有羽毛 THEN 该动物是鸟

r_4: IF 该动物会飞 AND 会下蛋 THEN 该动物是鸟

r_5: IF 该动物吃肉 THEN 该动物是食肉动物

r_6: IF 该动物有犬齿 AND 有爪 AND 眼盯前方
THEN 该动物是食肉动物

r_7: IF 该动物是哺乳动物 AND 有蹄
THEN 该动物是有蹄类动物

r_8: IF 该动物是哺乳动物 AND 是反刍动物
THEN 该动物是有蹄类动物

规则库

产生式系统例子

r_9： IF 该动物是哺乳动物 AND 是食肉动物 AND 是黄褐色
AND 身上有暗斑点 THEN 该动物是金钱豹

r_{10}： IF 该动物是哺乳动物 AND 是食肉动物 AND 是黄褐色
AND 身上有黑色条纹 THEN 该动物是虎

r_{11}： IF 该动物是有蹄类动物 AND 有长脖子 AND 有长腿
AND 身上有暗斑点 THEN 该动物是长颈鹿

r_{12}： IF 该动物是有蹄类动物 AND 身上有黑色条纹
THEN 该动物是斑马

r_{13}： IF 该动物是鸟 AND 有长脖子 AND 有长腿 AND 不会飞
AND 有黑白二色 THEN 该动物是鸵鸟

r_{14}： IF 该动物是鸟 AND 会游泳 AND 不会飞
AND 有黑白二色 THEN 该动物是企鹅

r_{15}： IF 该动物是鸟 AND 善飞 THEN 该动物是信天翁

产生式系统例子

- 设已知初始事实存放在综合数据库中：
  - 该动物身上有：暗斑点，长脖子，长腿，乳房，蹄

r_1：IF 该动物有毛发 THEN 该动物是哺乳动物

r_2：IF 该动物有乳房 THEN 该动物是哺乳动物

r_3：IF 该动物有羽毛 THEN 该动物是鸟

r_4：IF 该动物会飞 AND 会下蛋 THEN 该动物是鸟

r_5：IF 该动物吃肉 THEN 该动物是食肉动物

r_6：IF 该动物有犬齿 AND 有爪 AND 眼盯前方
THEN 该动物是食肉动物

r_7：IF 该动物是哺乳动物 AND 有蹄
THEN 该动物是有蹄类动物

r_8：IF 该动物是哺乳动物 AND 是反刍动物
THEN 该动物是有蹄类动物

产生式系统例子

- 设已知初始事实存放在综合数据库中：

该动物身上有：暗斑点，长脖子，长腿，乳房，蹄，哺乳动物

r_1：IF 该动物有毛发 THEN 该动物是哺乳动物

r_2：IF 该动物有乳房 THEN 该动物是哺乳动物

r_3：IF 该动物有羽毛 THEN 该动物是鸟

r_4：IF 该动物会飞 AND 会下蛋 THEN 该动物是鸟

r_5：IF 该动物吃肉 THEN 该动物是食肉动物

r_6：IF 该动物有犬齿 AND 有爪 AND 眼盯前方
THEN 该动物是食肉动物

r_7：IF 该动物是哺乳动物 AND 有蹄
THEN 该动物是有蹄类动物

r_8：IF 该动物是哺乳动物 AND 是反刍动物
THEN 该动物是有蹄类动物

√ →

“该动物是哺乳动物”

加入综合数据库

产生式系统例子

- 设已知初始事实存放在综合数据库中：

该动物身上有：暗斑点，长脖子，长腿，乳房，蹄，哺乳动物

r_1：IF 该动物有毛发 THEN 该动物是哺乳动物

r_2：IF 该动物有乳房 THEN 该动物是哺乳动物

√ →

“该动物是哺乳动物”

加入综合数据库

r_3：IF 该动物有羽毛 THEN 该动物是鸟

r_4：IF 该动物会飞 AND 会下蛋 THEN 该动物是鸟

r_5：IF 该动物吃肉 THEN 该动物是食肉动物

r_6：IF 该动物有犬齿 AND 有爪 AND 眼盯前方
THEN 该动物是食肉动物

r_7：IF 该动物是哺乳动物 AND 有蹄
THEN 该动物是有蹄类动物

√ →

“该动物是有蹄类动物”

加入综合数据库

r_8：IF 该动物是哺乳动物 AND 是反刍动物
THEN 该动物是有蹄类动物

产生式系统例子

- 设已知初始事实存放在综合数据库中：

该动物身上有：暗斑点，长脖子，长腿，乳房，蹄，哺乳动物，有蹄类动物

r_1： IF 该动物有毛发 THEN 该动物是哺乳动物

r_2： IF 该动物有乳房 THEN 该动物是哺乳动物

√ →

“该动物是哺乳动物”

加入综合数据库

r_3： IF 该动物有羽毛 THEN 该动物是鸟

r_4： IF 该动物会飞 AND 会下蛋 THEN 该动物是鸟

r_5： IF 该动物吃肉 THEN 该动物是食肉动物

r_6： IF 该动物有犬齿 AND 有爪 AND 眼盯前方

THEN 该动物是食肉动物

r_7： IF 该动物是哺乳动物 AND 有蹄

THEN 该动物是有蹄类动物

√ →

“该动物是有蹄类动物”

加入综合数据库

r_8： IF 该动物是哺乳动物 AND 是反刍动物

THEN 该动物是有蹄类动物

产生式系统例子

该动物身上有：暗斑点，长脖子，长腿，乳房，蹄，哺乳动物，有蹄类动物

r_9： IF 该动物是哺乳动物 AND 是食肉动物 AND 是黄褐色
AND 身上有暗斑点 THEN 该动物是金钱豹

r_{10}： IF 该动物是哺乳动物 AND 是食肉动物 AND 是黄褐色
AND 身上有黑色条纹 THEN 该动物是虎

r_{11}： IF 该动物是有蹄类动物 AND 有长脖子 AND 有长腿
AND 身上有暗斑点 THEN 该动物是长颈鹿

匹配成功！

长颈鹿

r_{12}： IF 该动物是有蹄类动物 AND 身上有黑色条纹
THEN 该动物是斑马

r_{13}： IF 该动物是鸟 AND 有长脖子 AND 有长腿 AND 不会飞
AND 有黑白二色 THEN 该动物是鸵鸟

r_{14}： IF 该动物是鸟 AND 会游泳 AND 不会飞
AND 有黑白二色 THEN 该动物是企鹅

r_{15}： IF 该动物是鸟 AND 善飞 THEN 该动物是信天翁

产生式系统的特点

- 不断从规则库中选择可用规则与综合数据库中的已知事实进行匹配
- 规则的每一次成功匹配都使综合数据库中增加新的内容，朝着问题解决的方向前进，称为推理，是专家系统中的核心内容

优点：

- 自然性
- 模块性
- 有效性
- 清晰性

缺点：

- 效率不高
- 不能表达结构性知识

适合产生式表示的知识：

- 领域知识间关系不密切，不存在结构关系，如化学方面。
- 经验性及不确定性的知识，且相关领域中对这些知识没有严格、统一的理论，如医疗论断。
- 领域问题的求解过程可被表示为一系列相对独立的操作，且每个操作可被表示为一条或多条产生式规则。

本章主要内容

1. 知识与知识表示的概念
2. 一阶谓词逻辑表示法
3. 产生式表示法
4. 框架表示法
5. 知识图谱

框架表示法

- 1975年，美国著名学者明斯基提出了框架理论：人们对现实世界中各种事物的认识都是以一种类似于框架的结构存储在记忆中的。

教室（名称）：有[属性]
墙，门，窗，天花板，
地板，课桌椅，讲台，
黑板……

- 框架表示法：一种结构化的知识表示方法，已在多种系统中得到应用。

框架的一般结构

- 框架（frame）是一种描述所论对象属性的数据结构
  - 一个框架由若干个被称为“槽”（slot）的结构组成，每一个槽又可根据实际情况划分为若干个“侧面”（facet）。
    - 一个槽用于描述所论对象某一方面的属性。
    - 一个侧面用于描述相应属性的一个方面。
    - 槽和侧面所具有的属性值分别被称为槽值和侧面值。

框架的一般结构

<框架名>

槽名1：　侧面名$_{11}$　　侧面值$_{111}$，…，侧面值$_{11P1}$

　　侧面名$_{1m}$　　侧面值$_{1m1}$，…，侧面值$_{1mPm}$

槽名n：　侧面名$_{n1}$　　侧面值$_{n11}$，…，侧面值$_{n1P1}$

　　侧面名$_{nm}$　　侧面值$_{nm1}$，…，侧面值$_{nmPm}$

约束：　约束条件$_{1}$

　　约束条件$_{n}$

用框架表示知识的例子

- 例1 教师框架

框架名：〈教师〉
姓名：单位（姓、名）
年龄：单位（岁）
性别：范围（男、女）
默认：男
职称：范围（教授，副教授，讲师，助教）
默认：讲师
部门：单位（系，教研室）
住址：〈住址框架〉
工资：〈工资框架〉
开始工作时间：单位（年、月）
截止时间：单位（年、月）
默认：现在

9个槽

用框架表示知识的例子

- 当把具体的信息填入槽或侧面后，就得到了相应框架的一个事例框架。

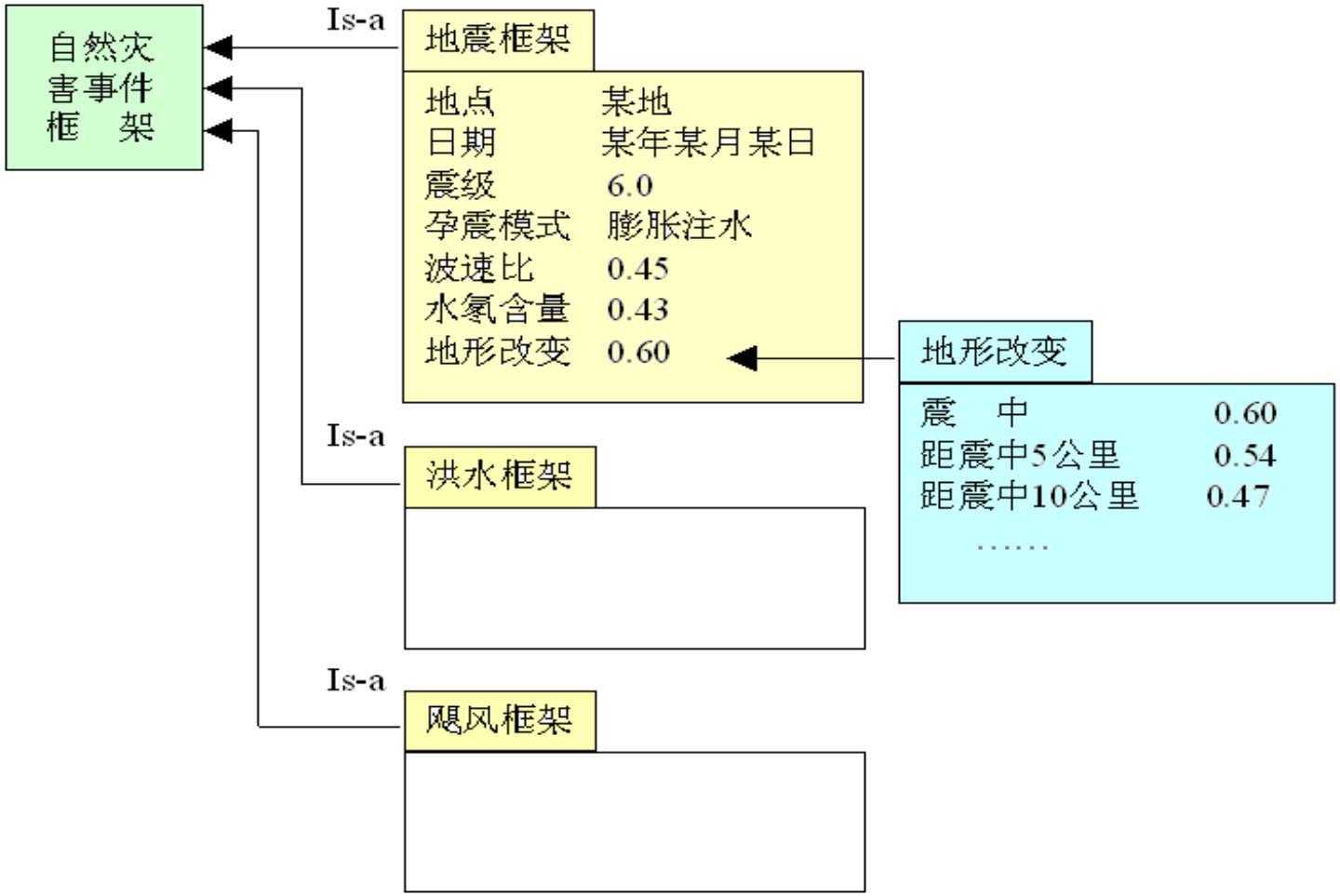
框架名：〈教师-1〉
姓名：张三
年龄：36
性别：女
职称：副教授
部门：人工智能与自动化学院
住址：〈adr-1〉
工资：〈sal-1〉
开始工作时间：1988.9
截止时间：1996.7

用框架表示知识的例子

- 例2 将下列一则地震消息用框架表示：“某年某月某日，某地发生6.0级地震，若以膨胀注水孕震模式为标准，则三项地震前兆中的波速比为0.45，水氡含量为0.43，地形改变为0.60。”

框架名：〈地震〉
地　　点：某地
日　　期：某年某月某日
震　　级：6.0
波 速 比：0.45
水氡含量：0.43
地形改变：0.60

用框架表示知识的例子


自然灾害事件框 架
Is-a
地震框架
地点 某地
日期 某年某月某日
震级 6.0
孕震模式 膨胀注水
波速比 0.45
水氢含量 0.43
地形改变 0.60
地形改变
震 中 0.60
距震中5公里 0.54
距震中10公里 0.47
.....
Is-a
洪水框架
Is-a
飓风框架

框架表示法的特点

（1） 结构性

- 便于表达结构性知识，能够将知识的内部结构关系及知识间的联系表示出来。

（2）继承性

- 框架网络中，下层框架可以继承上层框架的槽值，也可以进行补充和修改。

（3）自然性

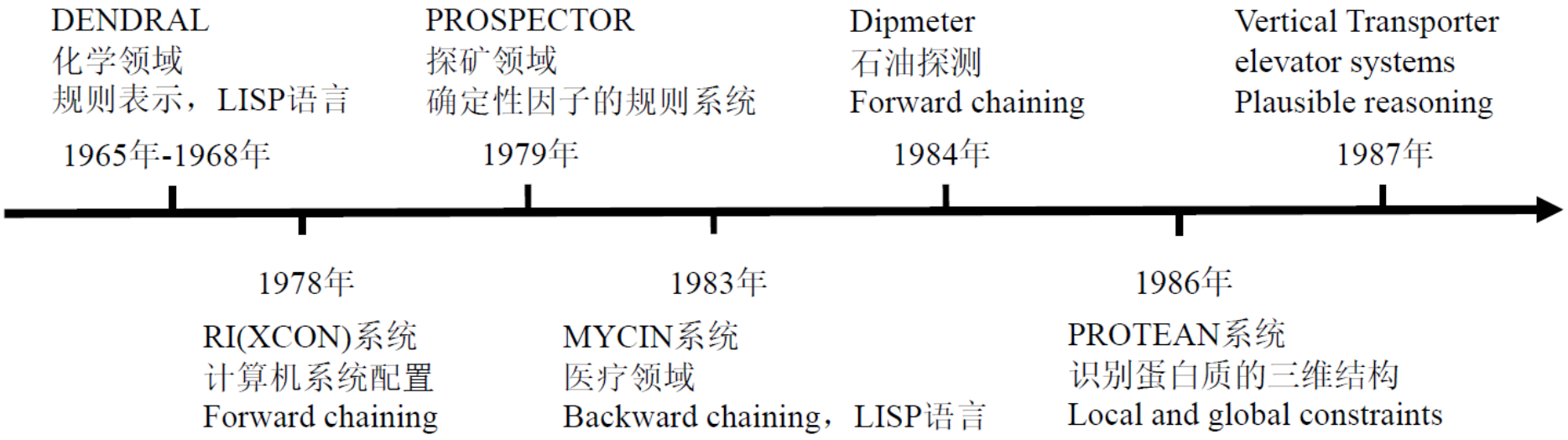
- 框架表示法与人在观察事物时的思维活动是一致的。

本章主要内容

1. 知识与知识表示的概念
2. 一阶谓词逻辑表示法
3. 产生式表示法
4. 框架表示法
5. 知识图谱

传统知识工程代表性系统

- 传统知识工程在规则明确、边界清晰、应用封闭的应用场景取得了巨大成功



传统方法的特点和困难

- 自上而下：严重依赖专家和人的干预
- 知识获取困难：隐性知识、过程知识等难以表达，存在主观性、不一致性，难以完备
- 知识应用困难：超出知识边界、需要常识、处理异常、不确定性推理、知识更新

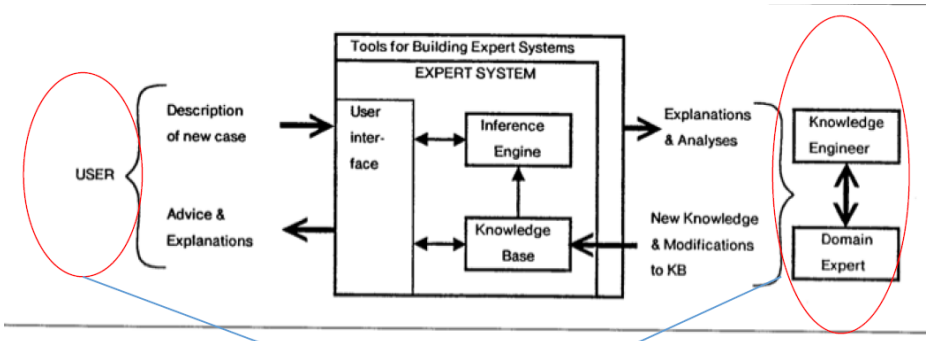
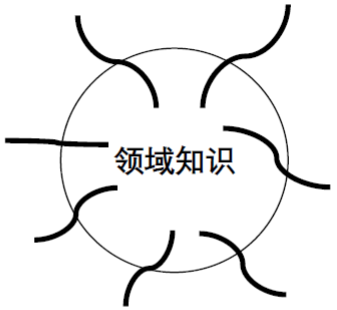


FIGURE 1-2 Interaction of a knowledge engineer and domain expert with software tools that aid in building an expert system. Arrows indicate information flow.

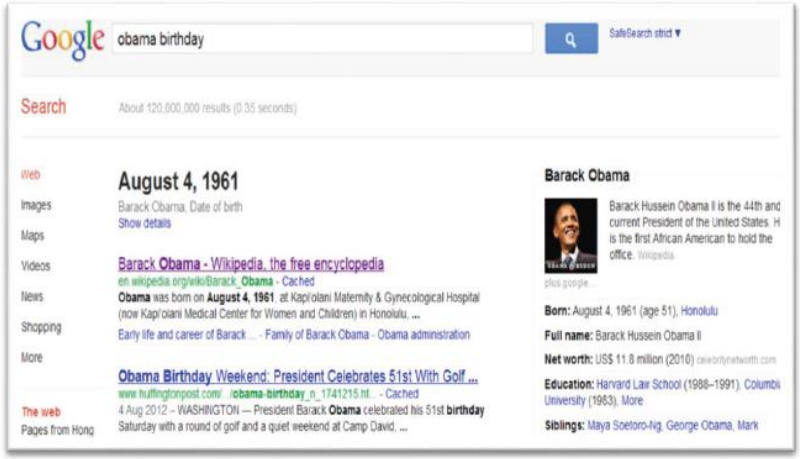
MYCIN专家系统中的人工参与部分



开放知识

知识图谱的诞生

- 2012年5月，Google收购Metaweb公司，并正式发布知识图谱
- 搜索核心需求： 让搜索通往答案
  - 无法理解搜索关键词
  - 无法精准回答
- 根本问题
  - 缺乏大规模背景知识
  - 传统知识表示难以满足需求



知识图谱的诞生

- Google介绍Knowledge Graph

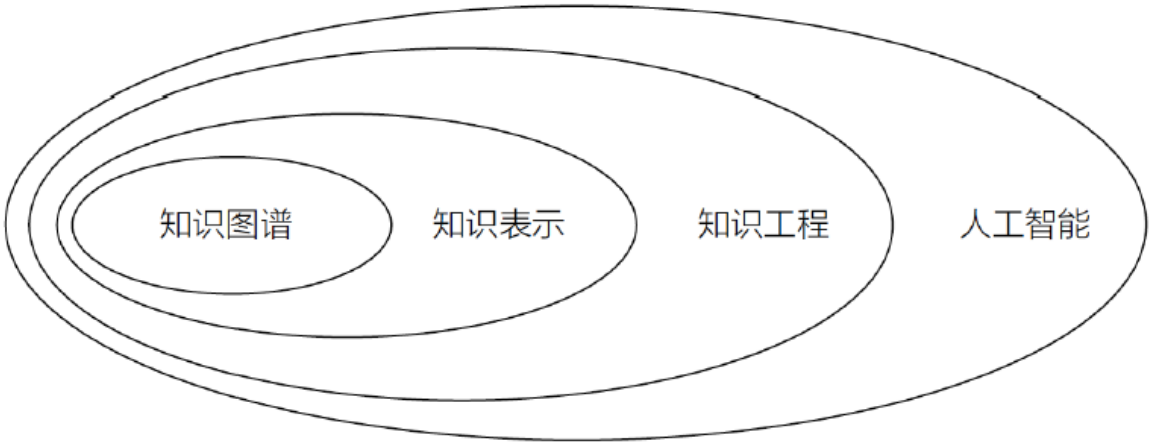
https://www.bilibili.com/video/BV1JJ411T7zg?from=search&seid=14222123860998730516&spm_id_from=333.337.0.0
B站搜：谷歌官方 知识图谱介绍

或

https://www.youtube.com/watch?v=mmQl6VGvX-c

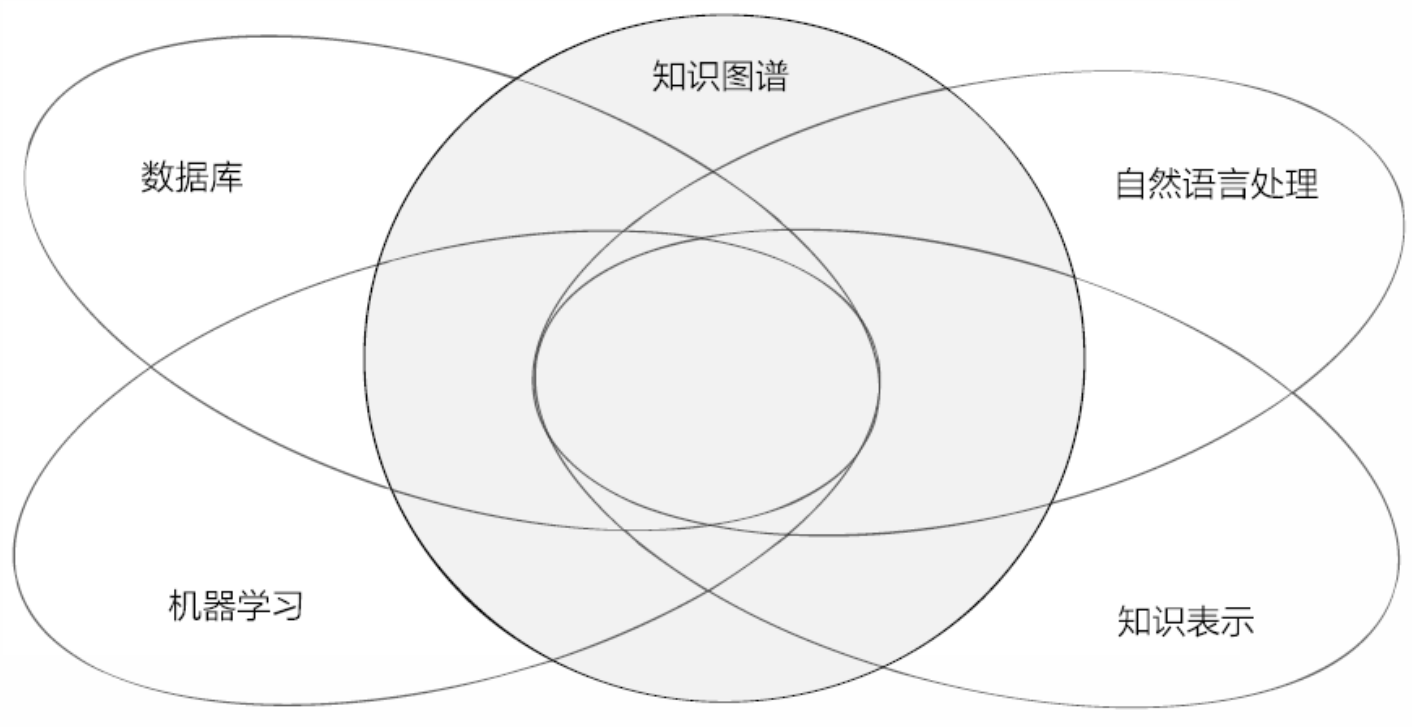
知识图谱与人工智能

- 作为一种技术体系，是大数据时代知识工程的代表性进展
- 作为一门学科，知识图谱属于人工智能范畴
- 知识表示是发展知识工程最关键的问题之一，而知识表示的一个重要方式就是知识图谱



知识图谱的学科地位

知识图谱的广义内涵



知识图谱的概念

- 知识图谱(Knowledge Graph)本质上是一种大规模语义网络（semantic network）
  - 富含实体(entity)、概念(concepts)及其之间的各种语义关系(semantic relationships)
  - 是大数据时代知识表示的重要方式之一

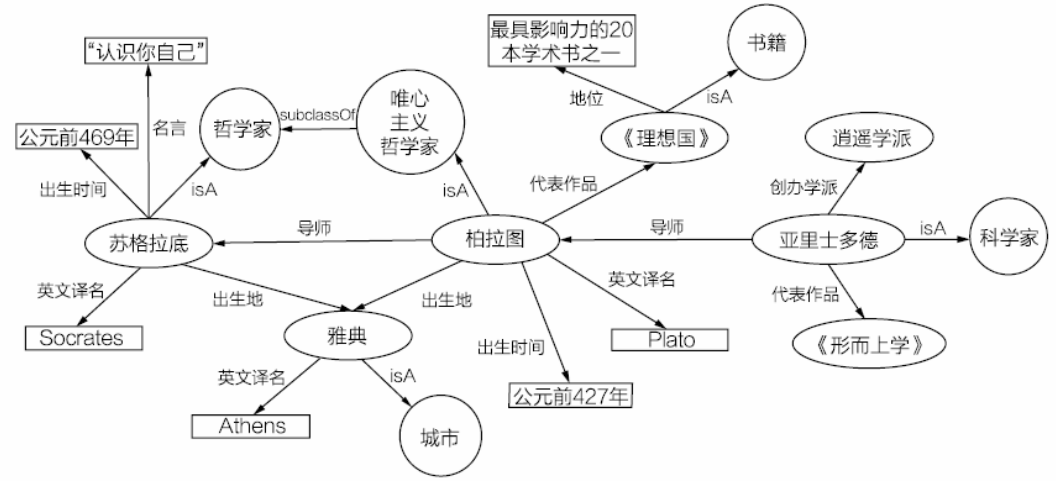
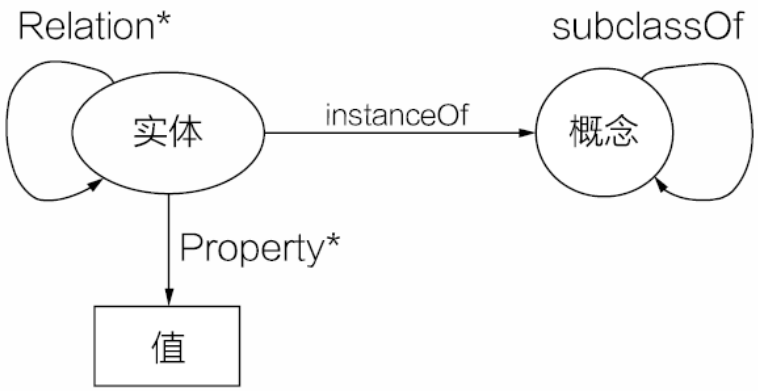


图 1-1　关于古希腊三大哲学家的知识图谱片段

语义网络

- 语义网络是一种以图形化的(Graphic)形式通过点和边表达知识的方式，其基本组成元素是点和边
- 1968年罗斯 奎利恩最先提出



语义网络的组成（图中星号表示可以存在多个不同的属性或者关系）

知识图谱的组成：节点

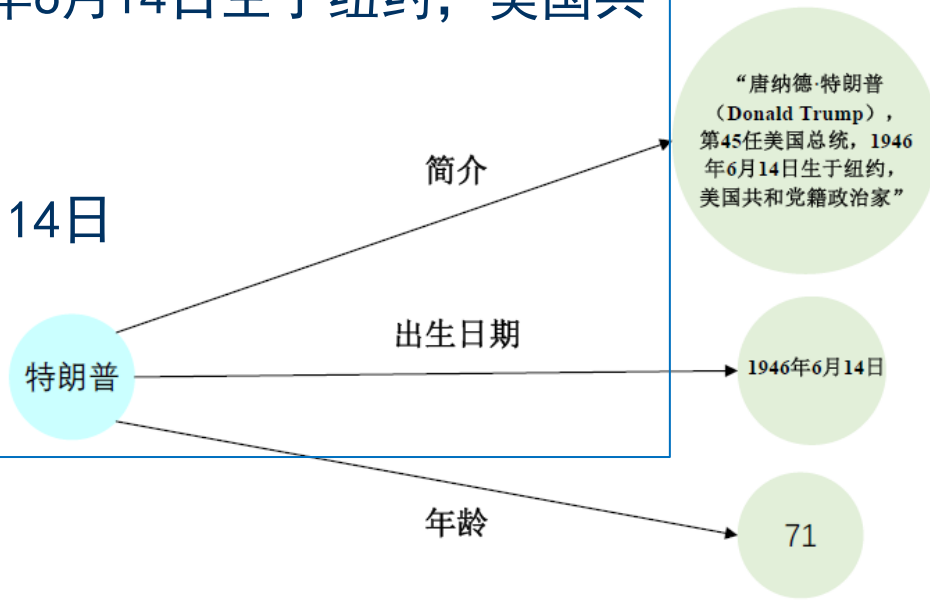
- 实体 Entity/Objects/Instances
  - 具有可区别性且独立存在的某种事物
    - Wikipedia: An entity is something that exists as itself, as a subject or as an object, actually or potentially, concretely or abstractly, physically or not.
    - 黑格尔《小逻辑》：能够独立存在的，作为一切属性的基础和万物本原的东西
- 概念 Concept/Category/Type/Class
  - 具有同种特性的实体构成的集合
    - Concept: In metaphysics, and especially ontology, a concept is a fundamental category of existence.
    - (mental) representations of categories
    - Category: Groups of entities which have something in common
    - Type/Class: A grouping based on shared characteristics.

知识图谱的组成：节点

- 值 Value
  - 对象指定属性的值

例：

- String
  - 特朗普简介：“唐纳德·特朗普（Donald Trump），第45任美国总统，1946年6月14日生于纽约，美国共和党籍政治家”
- Date
  - 特朗普生日：1946年6月14日
- Numeric
  - 特朗普年龄71



知识图谱的组成：边

- 关系 Relation
  - 侧重实体之间的关系
    - Examples:

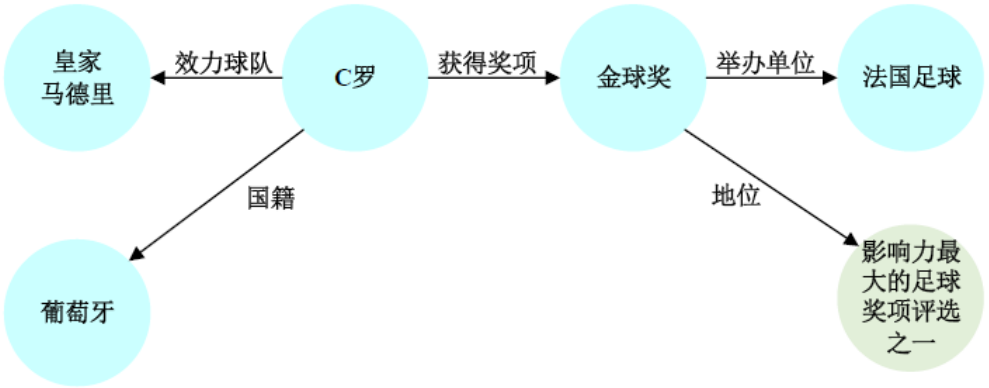
      Sitting-On: An apple sitting on a table

      Taller-than: Washington Monument is taller than the White House

- 属性 Property/Attribute/Quality
  - 描述一个物体的特性
    - Examples:

      size, color, weight,

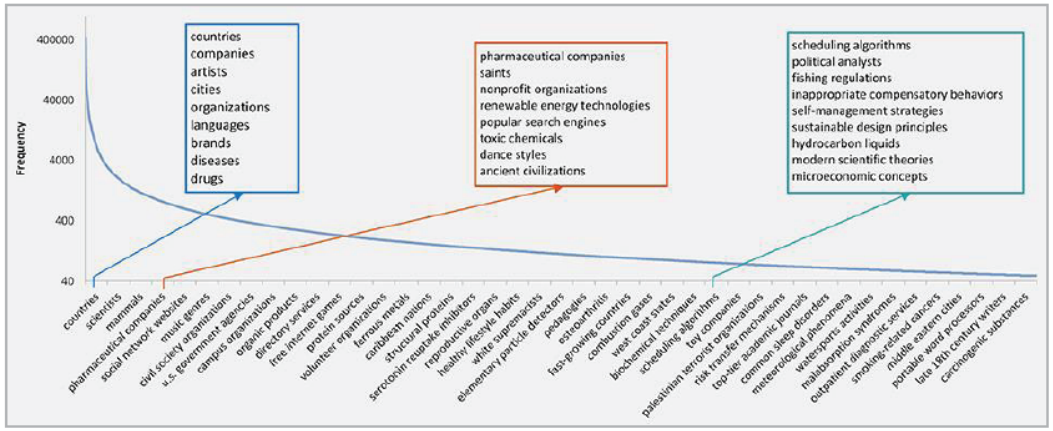
      composition of an object



知识图谱的优势

- 尺度大 large scale
  - Higher coverage over entities and concepts

KGs	# of Entities/Concepts	# of Relations
YAGO	10 Million	120 Million
DBpedia	28 Million	9.5 **Billion**
Probase	2.7 Million	70 **Billion**
BabelNet	14 Million	**5 Billion**
CN-DBpedia	17 Million	200 Million

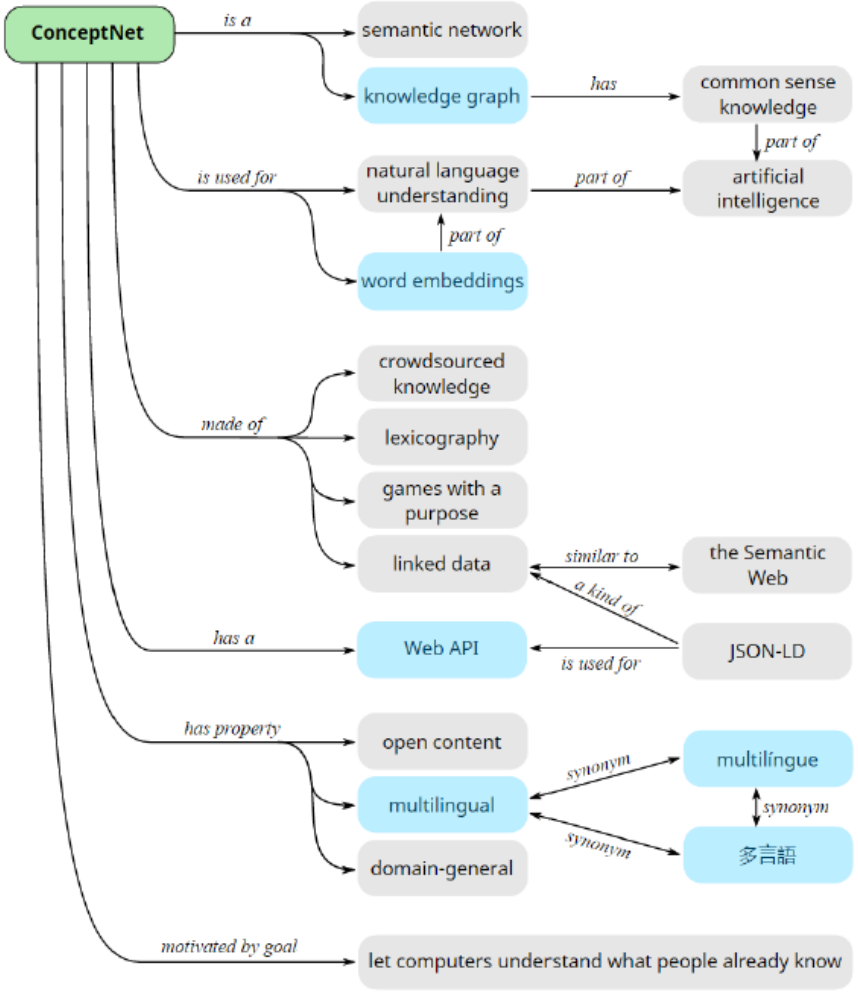


Existing Taxonomies	Number of Concepts
Freebase [5]	1,450
WordNet [13]	25,229
WikiTaxonomy [26]	111,654
YAGO [35]	352,297
DBPedia [1]	259
ResearchCyc [18]	≈ 120,000
KnowItAll [12]	N/A
TextRunner [2]	N/A
OMCS [31]	N/A
NELL [7]	123
Probase	**2,653,872**

知识图谱的优势

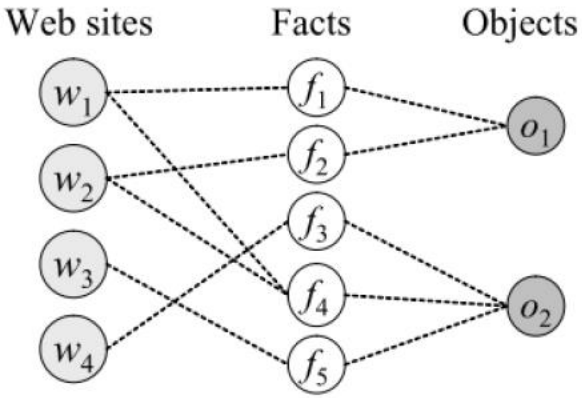
- 语义丰富 semantically rich
  - Higher coverage over numerous semantic relationships

KGs	# of Relations
DBpedia	1,650
YAGO1	14
YAGO3	74
CN-DBpedia	100 Thousands



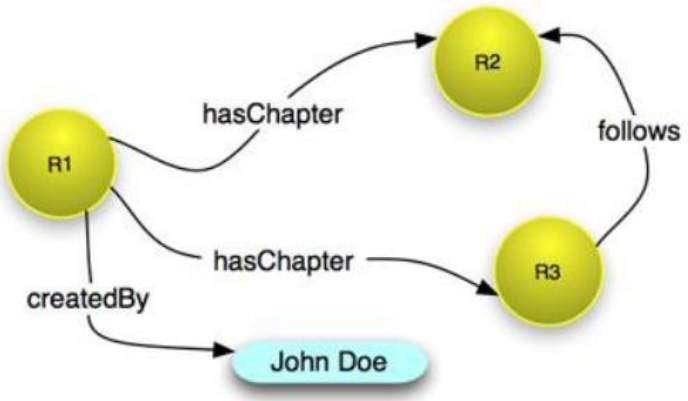
知识图谱的优势

- 质量高 high quality
  - 大数据 Big data: Cross validation by multiple sources
  - 众包 Crowd sourcing: quality guarantee

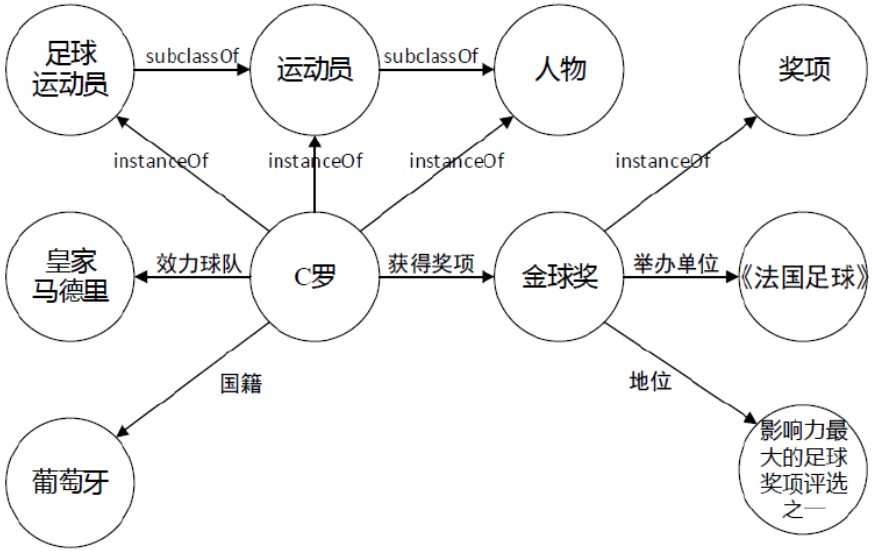


知识图谱的优势

- 结构易于实现 friendly structure
  - 组织架构
    - By RDF
    - By graph



Subject	Predicate	Object
R1	hasChapter	R2
R1	hasChapter	R3
R3	follows	R2
R1	createdBy	"John Doe"

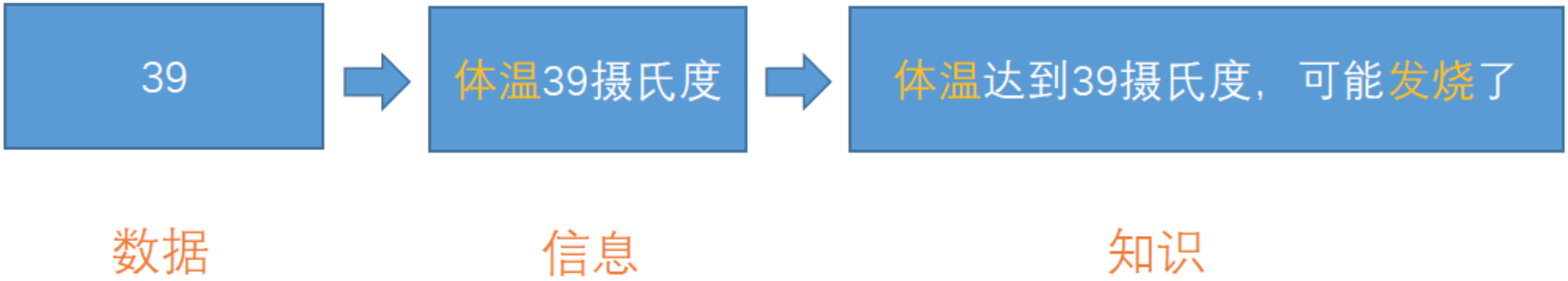


知识图谱的挑战

- 高质量模式缺失
  - 知识图谱在设计模式时通常会采取一种“经济、务实”的做法：也就是允许模式（Schema）定义不完善，甚至缺失
  - 模式定义不完善或缺失对知识图谱中的数据语义理解以及数据质量控制提出了挑战
- 封闭世界假设不再成立
  - 传统数据库与知识库的应用通常建立在封闭世界假设（CWA）基础之上，大多数开放性应用不遵守这一假设，在这些应用中缺失的事实或知识未必为假
  - 不遵守CWA给知识图谱上的应用带来了巨大的挑战
- 大规模自动化知识获取成为前提
  - 大规模自动化知识获取是知识图谱与传统语义网络的根本区别

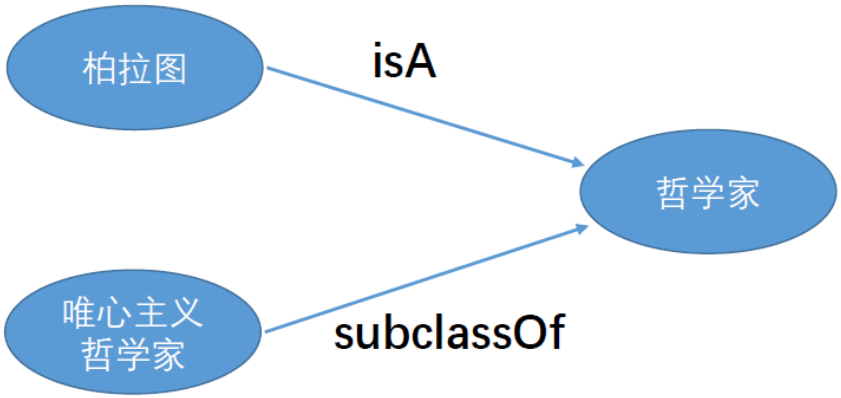
数据、信息与知识

- 数据：对客观世界的符号化记录
- 信息：被赋予意义的数据
- 知识：信息之间有意义的关联



知识的类别

- 事实知识：关于某个特定实体的基本事实
- 概念知识：实体与概念之间的类属关系/子概念与父概念之间的子类关系



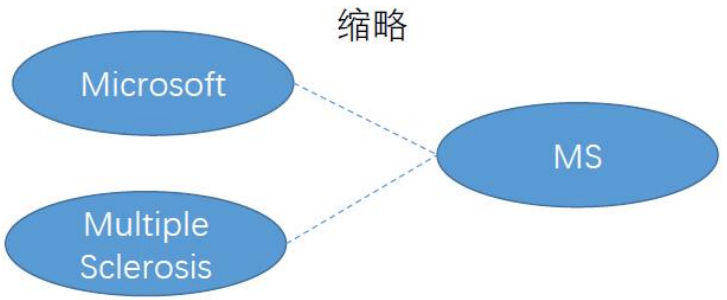
yago
select knowledge

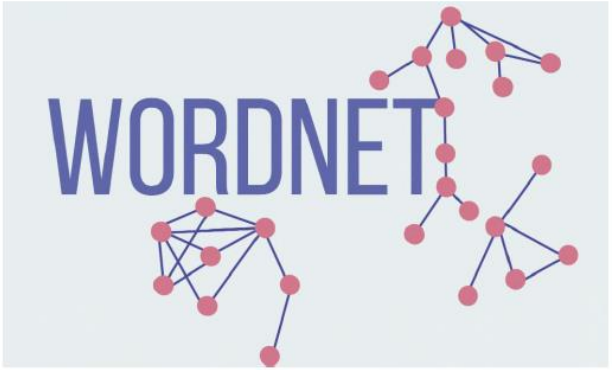
ProBase

典型概念知识图谱

知识的类别

- 词汇知识：实体与词汇/词汇与词汇
- 常识知识：常识属性/常识规则
- 其他知识：上下文知识、相关联知识、空间知识……

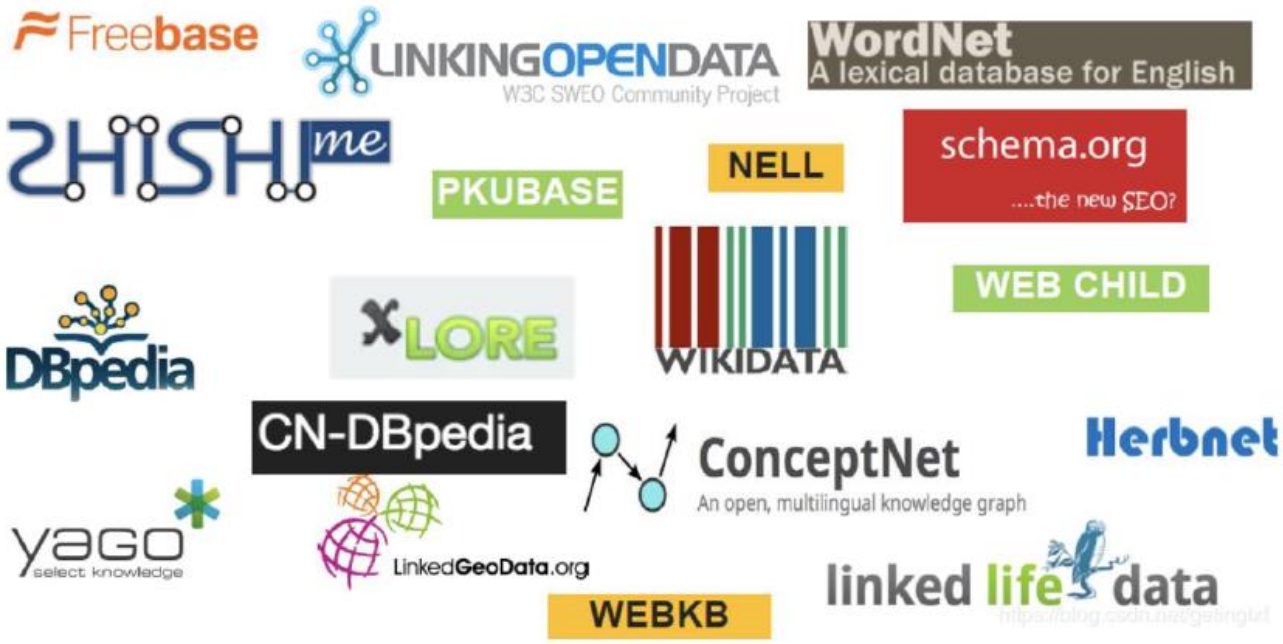




典型词汇知识图谱

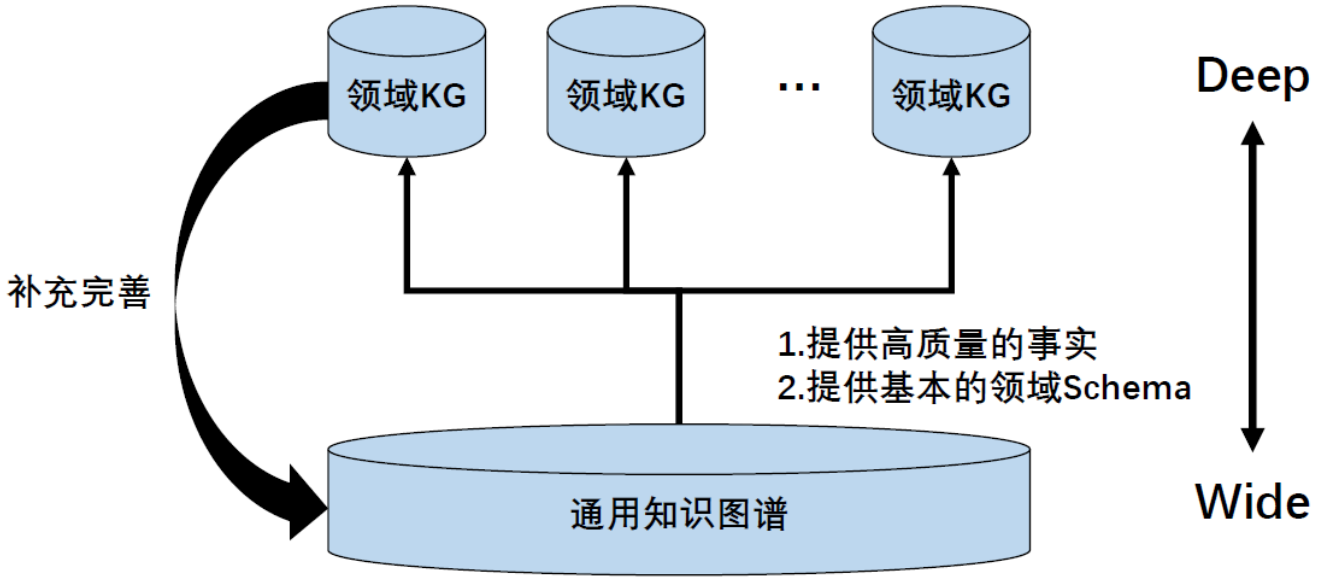
针对不同种类知识的知识图谱

- 概念图谱
- 常识图谱
- 词汇图谱
- 百科图谱
- 文本图谱
- ……



知识图谱的领域特性

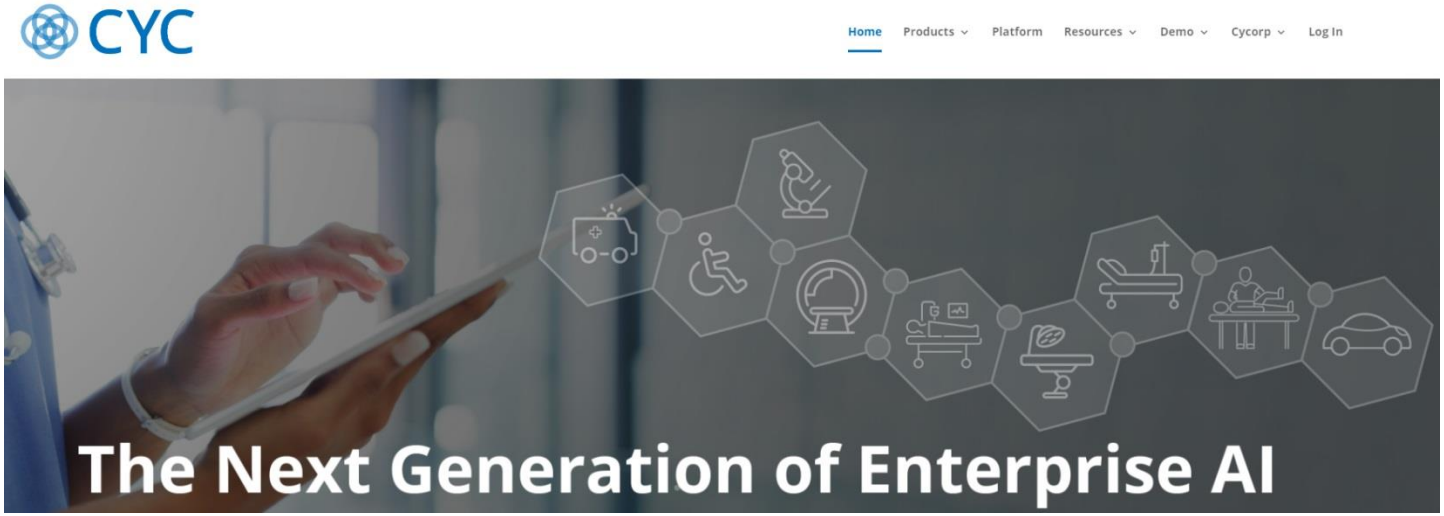
- 通用知识图谱（General-purpose Knowledge Graph）
- 领域（行业）知识图谱（Domain-specific Knowledge Graph）
- 企业知识图谱（Enterprise Knowledge Graph）



通用知识图谱与领域知识图谱的关系

常见知识图谱

- Cyc——常识知识图谱
  - 通过人工方法将上百万条人类常识编码成机器可用的形式，用以进行智能推断
  - 目前ResearchCyc知识图谱中包含了700万条断言（事实和规则），涉及63万个概念，38000种关系



https://cyc.com/

常见知识图谱

- WordNet——基于认知语言学的英语词典
  - 以同义词集合（synset）作为一个基本单元
  - 规模：

POS	*Unique Strings*	*Synsets*	*Total Word-Sense Pairs*
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

https://wordnet.princeton.edu/

常见知识图谱

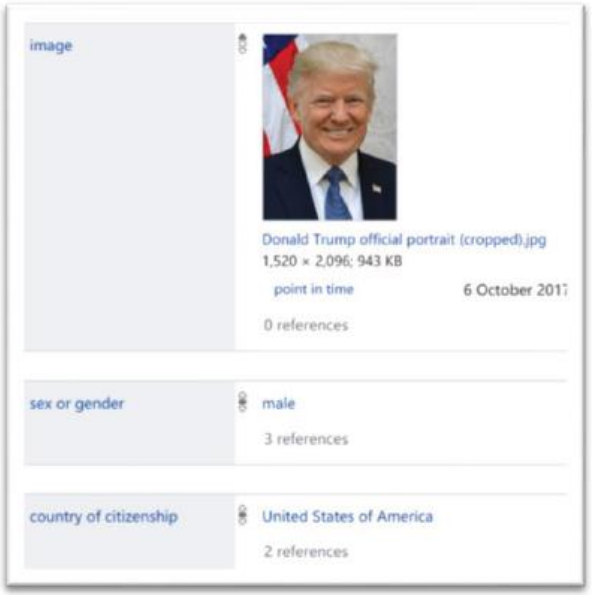
- ConceptNet——大型的多语言常识知识库
  - 知识来源丰富
    - 众包(Crowd-Sourcing)
    - 资源（例如Wiktionary 和Open Mind Common Sense）
    - 带目的的游戏（如Verbosity和nadya.jp）
    - 专家创建的资源(如WordNet和JMDict)

http://conceptnet.io/

常见知识图谱

- Freebase/Wikidata
  - Freebase所有知识采用结构化的表示形式，可由机器和人编辑
  - Wikidata是维基百科的姐妹工程，同样可由机器和人自由编辑
  - 2016年8月31日，Freebase宣布关闭，所有数据汇入Wikidata

  - 众包构建
  - 结构化三元组
  - Wikidata目前包含49,915,906个实体



常见知识图谱

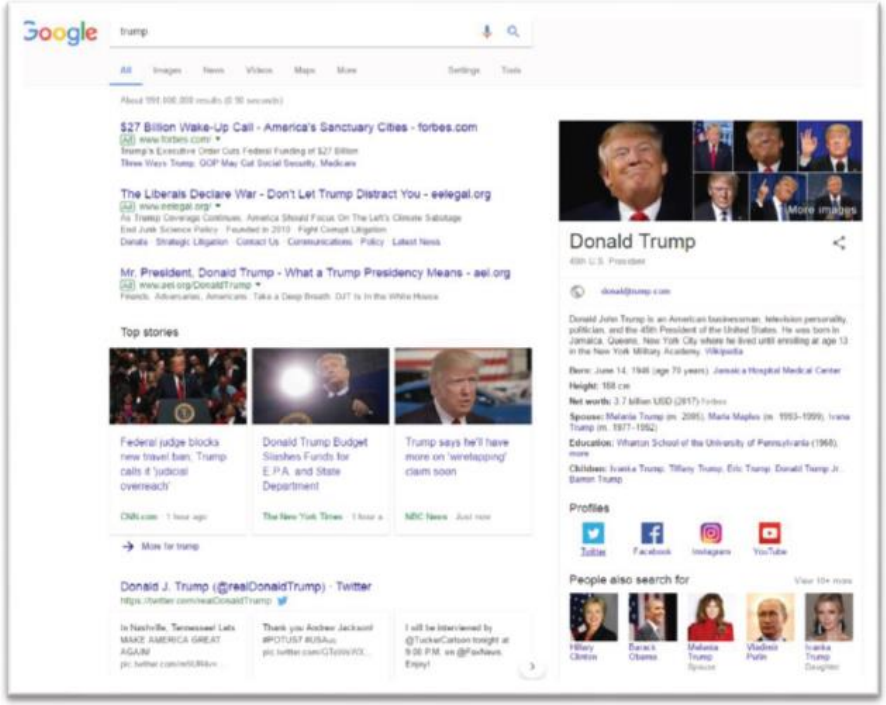
- Dbpedia
  - 从维基百科页面中自动抽取出结构化的知识，构建而成的大型通用百科图谱
  - 多语言
  - 自动构建
  - 共收录有127种不同语言共计2800万实体，其中英文实体数量最大，为467万

```
<http://dbpedia.org/resource/A> <http://dbpedia.org/property/name> "Latin Capital Letter A"@en .
<http://dbpedia.org/resource/A> <http://dbpedia.org/property/name> "Latin Small Letter A"@en .
<http://dbpedia.org/resource/A> <http://dbpedia.org/property/map> "ASCII 1"@en .
```

http://wiki.dbpedia.org/

常见知识图谱

- Google KG
  - 谷歌知识图谱于2012年发布，被认为是搜索引擎的一次重大革新
  - 规模巨大
  - 用于增强搜索引擎的搜索能力
  - 5700万实体，180亿关系



常见知识图谱

- Probase
  - 概念图谱
  - 数据源来自微软搜索引擎Bing的网页，主要利用Hearst Pattern从文本中抽取IsA关系
  - 概念规模最大
  - 自动构建
  - 1200万实体，540万概念

From: "... in **tropical countries** such as **Singapore**, **Malaysia**, ..."
To:

- (Singapore, isA, tropical countries)
- (Malaysia, isA, tropical countries)

常见知识图谱

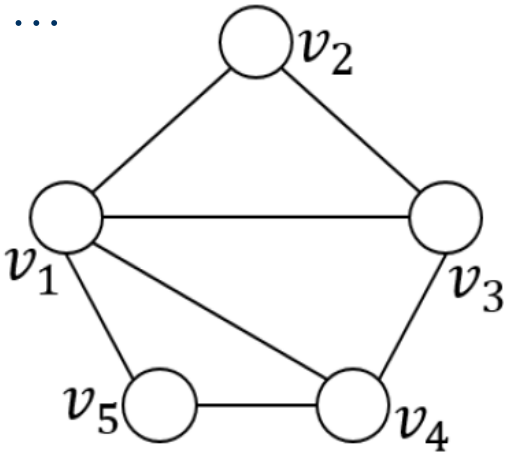
- 搜狗知立方
  - 中文知识图谱，应用于搜狗搜索引擎
  - 侧重于娱乐领域

- 百度知心
  - 中文知识图谱，应用于百度搜索引擎
  - 融合百度百科知识

基于图论的知识图谱表示

- ***G=G(V, E)***
  - 其中V表示顶点集，$E \subseteq V \times V$表示边的集合。

  - 有向图、无向图
  - 邻接表、邻接矩阵
  - 度数、路径、可达 …



0	1	1	1	1
1	0	1	0	0
1	1	0	1	0
1	0	1	0	1
1	0	0	1	0

基于三元组的知识图谱表示

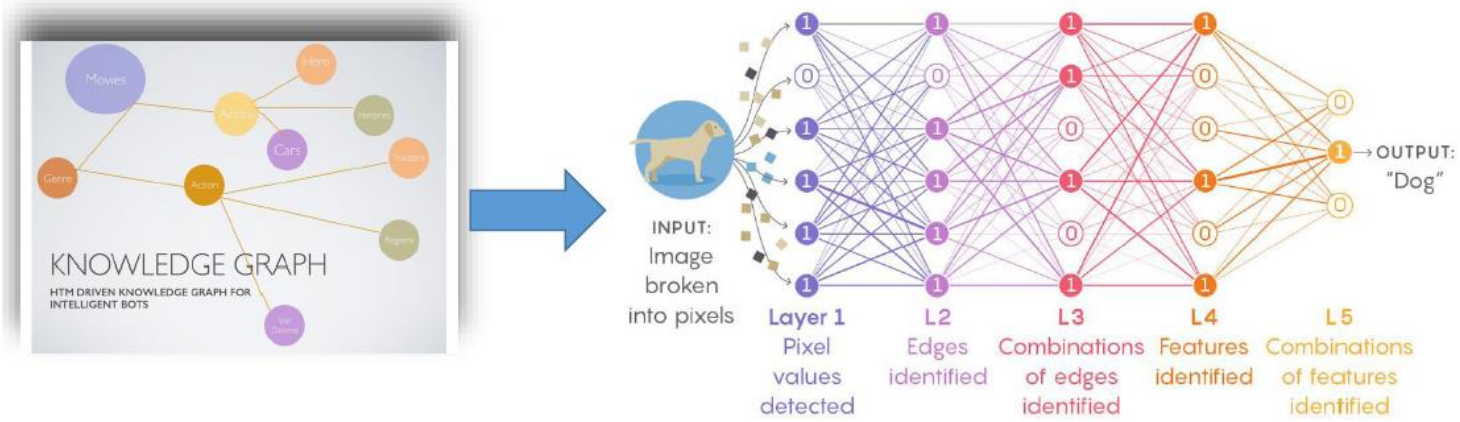
- RDF（Resource Description Framework）
  - 是用于描述现实中资源的W3C标准。
  - 现实中任何实体都可以表示成RDF模型中的资源，这些资源是对现实世界中概念、实体和事件的抽象。
- 三元组包括三个元素：主体（subject)、属性（property）及客体(object)
  - 也被称为主体、属性及属性值（property value）

e. g. 〈亚理士多德，受到影响，柏拉图〉

主体（Subject）	谓词（Predicate）	客体（Object）
Aristotle	*influencedBy*	*Plato*
Boethius	*placeOfDeath*	*Pavia*
Chalcis	*country*	*Greece*
Pavia	*postalCode*	27100

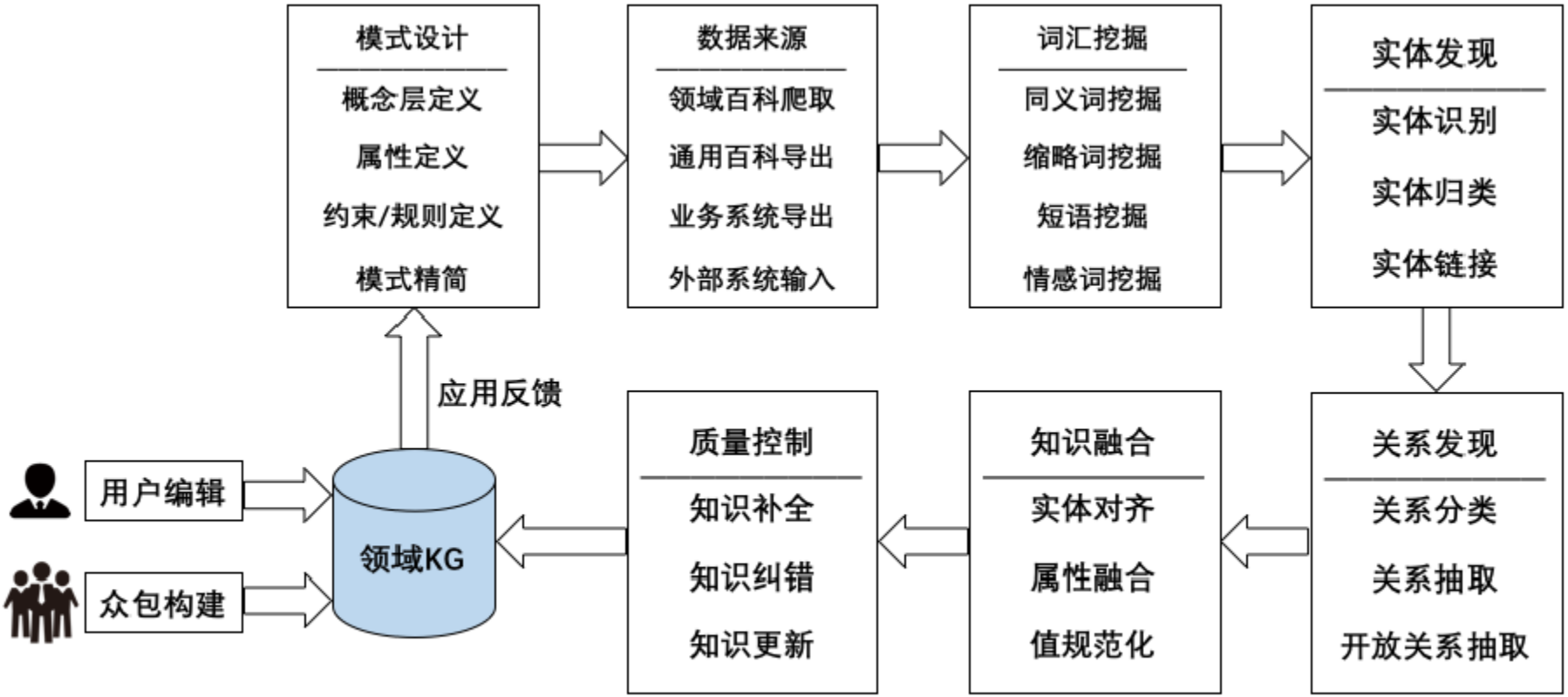
基于数值的知识图谱表示

- 将知识图谱中元素(包括实体、属性概念等)表示为低维稠密实值向量。
- 知识图谱的不同表示各有适用场景：
  - 向量化的表示面向机器处理
  - 符号化表示面向人的理解
  - 符号表示更易于理解、实现符号推理



将知识图谱中的点与边表达成数值化向量

领域知识图谱构建



领域知识图谱构建的基本流程

领域知识图谱构建

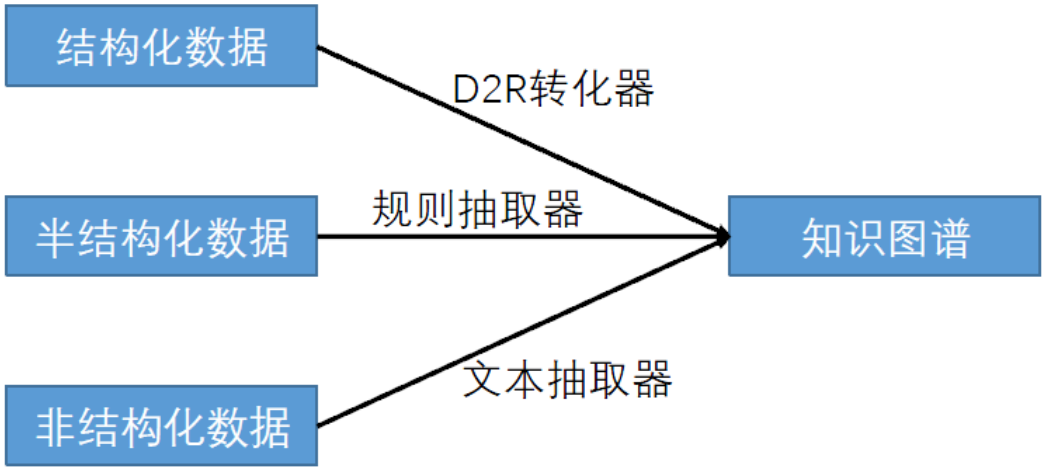
- **1. 模式设计**
  - 将认知领域的基本框架赋予机器
  - 概念层设计
    - 指定领域的基本概念，以及概念之间子类关系

    e.g., 足球领域，足球运动员是运动员的子类
  - 属性定义
    - 明确领域的基本属性，明确属性的适用概念，属性值的类别或范围

    e.g.,效力球队的域为足球运动员，范围为球队
  - 约束规则定义
    - 多值属性约束 e.g., 出生日期（单值），获得奖项（多值）
    - 互逆属性约束 e.g., 隶属球员和效力球队为互逆属性

领域知识图谱构建

- **2. 明确数据来源**
  - 结构化程度较高、质量较好，以尽可能低代价获取数据
    - 互联网上的领域百科爬取
    - 通用百科图谱的导出
    - 内部业务数据的转换
    - 外部业务系统的导入



不同数据来源通过不同的知识获取方式构建知识图谱

领域知识图谱构建

- **3. 词汇挖掘**
  - 识别出领域中重要短语和词汇
    - 识别领域的高质量词汇
    - 识别同义词
    - 识别缩写词
    - 识别领域常见情感词

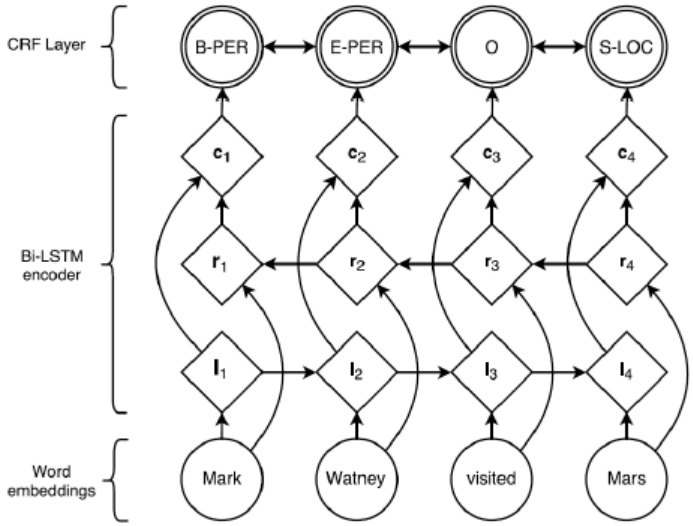
Raw Corpus

Quality Phrases

Jiawei Han, etc.,. Mining Quality Phrases from Massive Text Corpora

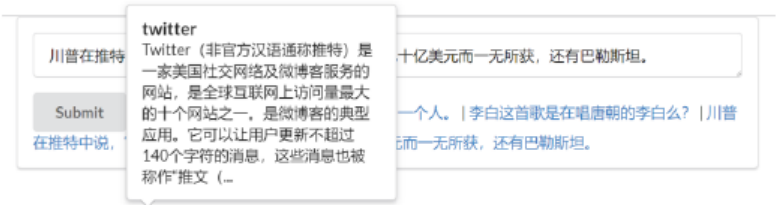
领域知识图谱构建

- **4. 实体发现**
  - 识别出领域中的常见实体
  - 理解领域文本和数据的关键一步
    - 实体识别
    - 实体归类
    - 实体链接



Guillaume Lample etc., Neural Architectures for Named Entity Recognition

实体链接



知识工场实验室的实体链接DEMO

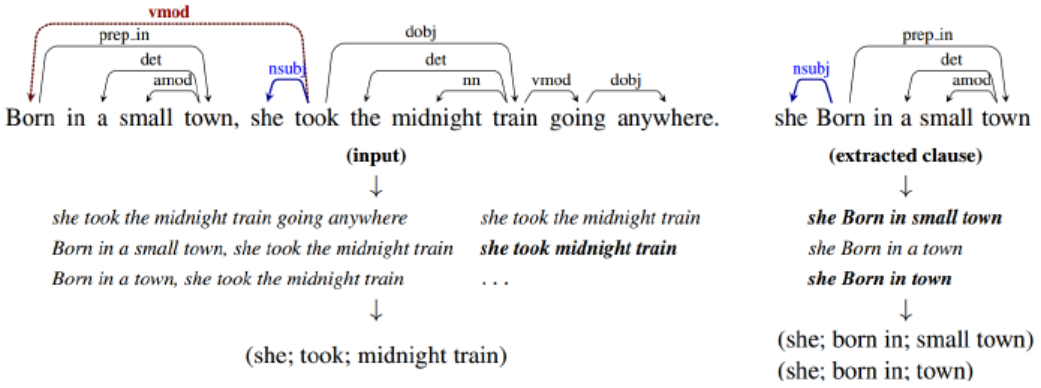
领域知识图谱构建

- **5. 关系发现**
  - 填充知识库中的关系实例
    - 关系分类：将给定的实体对（entity pairs）分类到某个已知关系

    e.g., 李娜-姜山 ->丈夫，教练
    - 关系抽取：从文本中抽取某个实体对的具体关系

    e.g., 姜山曾先后两次成为李娜的教练->(李娜，教练，姜山）
    - 开放关系抽取：从文本中抽取出实体对之间的关系描述

    e.g., 上海隔中国东海与日本九州岛相望->(上海，相望，日本九州岛）



Stanford Open Information Extraction,
https://nlp.stanford.edu/software/openie.html

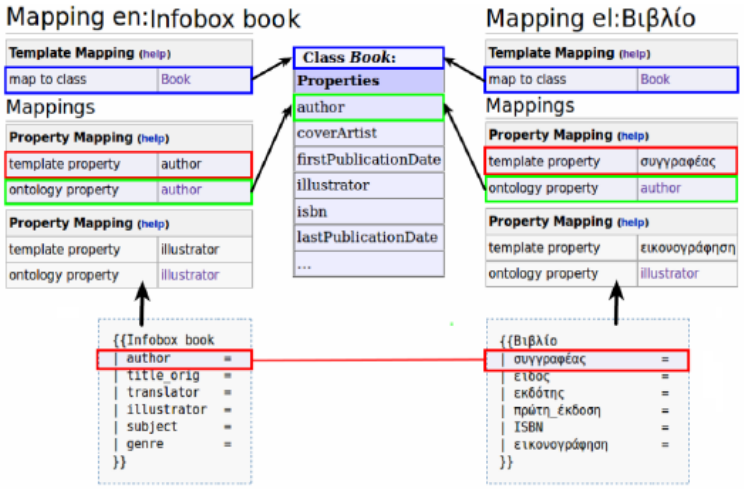
领域知识图谱构建

- **6. 词汇挖掘**
  - 融合来自不同数据源的知识
    - ➢ 实体对齐：识别不同来源的统一实体

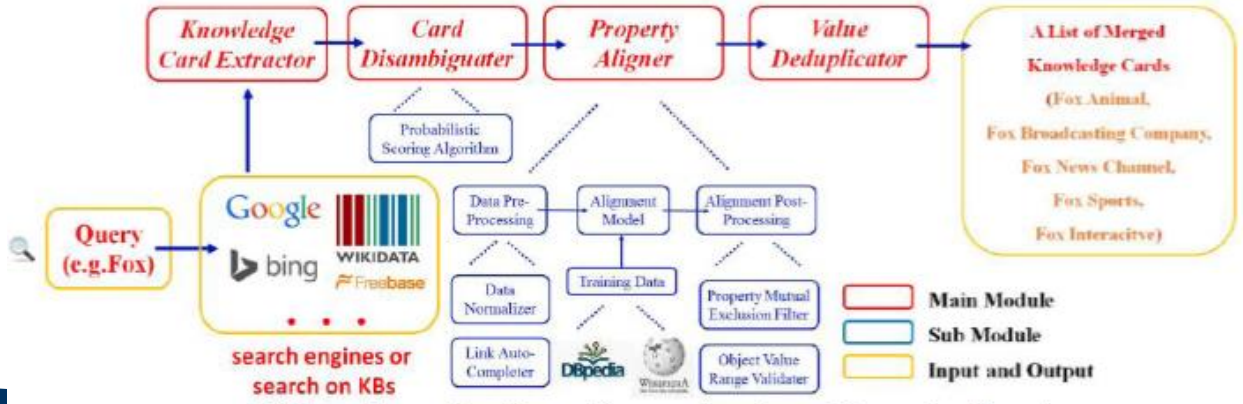
      e.g., 华中科技大学，HUST
    - ➢ 属性融合：识别同一属性的不同描述

      e.g.,英文名，英文名称
    - ➢ 值规范化：规范化到统一格式/单位

      e.g., 175cm，1米75



跨语言知识融合



Effective Online Knowledge Graph Fusion

领域知识图谱构建

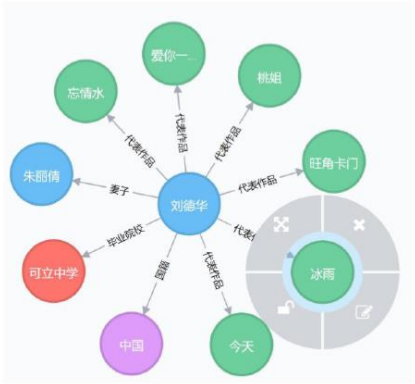
- **7. 质量控制**
  - ➢ 知识补全

  e.g., 如果一个人出生地是中国，推断其国籍也可能是中国

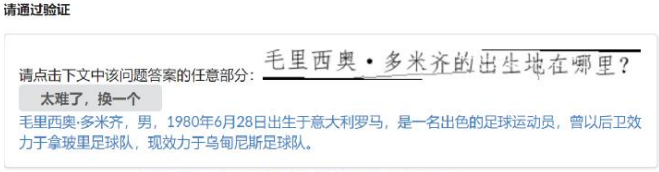
  e.g., 从外部互联网文本数据补充知识
  - ➢ 知识纠错

  e.g., 互逆属性纠错：A妻子B，B丈夫C
  - ➢ 知识更新
- **8. 人工干预**
  - ➢ 人工编辑
  - ➢ 众包构建

  e.g., 利用知识问答验证码来进行知识获取



知识工场实验室推出的KADE系统，能够所见即所得的知识图谱编辑



基于文本理解的超级验证码可以实现大规模众包化知识获取

提升知识图谱的质量

知识图谱的应用