知识图谱（Knowledge Graph, KG）

- 知识图谱(Knowledge Graph)以结构化的方式描述客观世界中概念、实体及其之间的关系；
- 本质上是一种大规模语义网络(semantic network)

- 知识图谱通过对错综复杂的数据进行有效的加工、处理、整合，转化为简单、清晰的“实体-关系-实体”的三元组，最后聚合大量知识；

知识表示的重要性

- 知识是智能的基础
  - 机器可以获得知识
  - 机器可以运用知识
- 符合计算机要求的知识模式
  - 计算机能存储、处理的知识表示模式
  - 数据结构（List, Table, Tree, Graph, etc.）

语义网：

RDF模型

- 在RDF中，知识总是以三元组形式出现
- 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))

知识图谱的分布式表示

在保留语义的同时，将知识图谱中的实体和关系映射到连续的稠密的低维向量空间

知识抽取

从不同来源、不同结构的数据中进行知识提取，形成知识存入到知识图谱。

什么是关系抽取？

- 信息抽取（Information Extraction）研究领域的任务之一
- 从文本中抽取出两个或者多个实体之间的语义关系

、

知识图谱优势：规模大，语义丰富，质量高，结构化

知识获取关键技术与难点

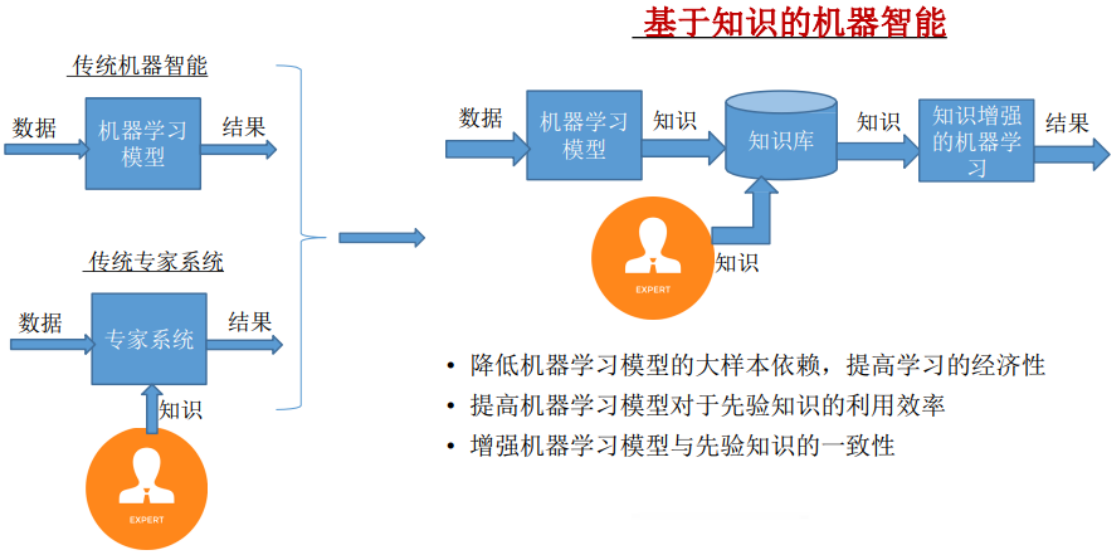
- 从结构化数据库中获取知识：D2R
  - 难点：复杂表数据的处理
- 从链接数据中获取知识：图映射
  - 难点：数据对齐
- 从半结构化（网站）数据中获取知识：使用包装器
  - 难点：方便的包装器定义方法，包装器自动生成、更新与维护
- 从文本中获取知识：信息抽取
  - 难点：结果的准确率与覆盖率

认知智能是智能化的关键

Can machine **think like humans**?

- 理解与解释是后深度学习时代人工智能的核心使命之一

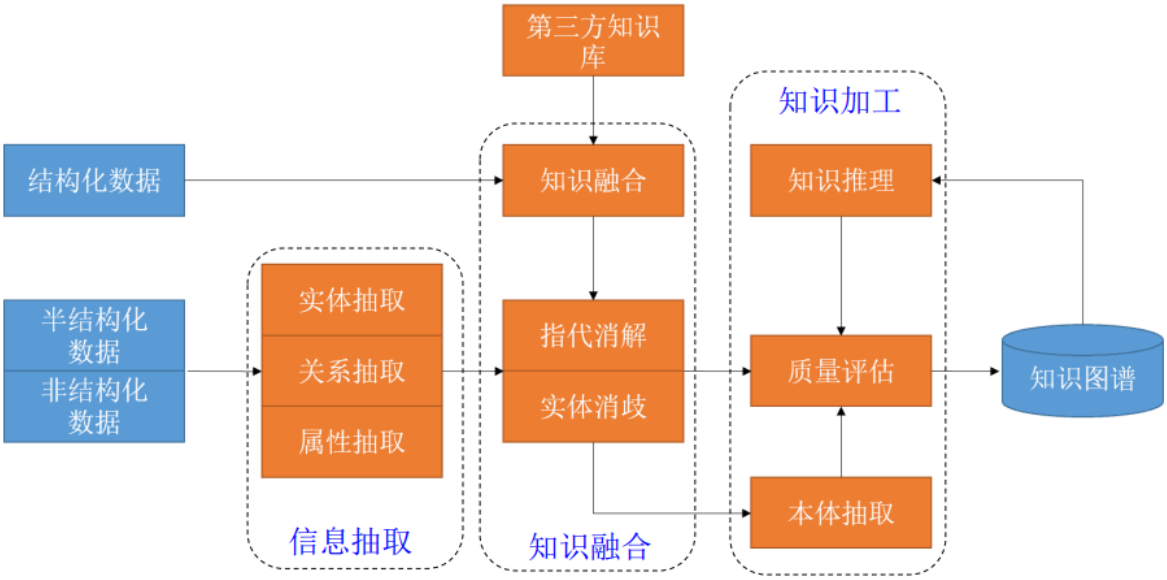
知识增强机器学习能力



- 降低机器学习模型的大样本依赖，提高学习的经济性
- 提高机器学习模型对于先验知识的利用效率
- 增强机器学习模型与先验知识的一致性

知识图谱的构建

知识图谱的技术架构：



16

知识图谱的构建

知识图谱的技术架构：

•**信息抽取**：从各种类型的数据源中提取出实体、属性以及实体间的相互关系，在此基础上形成本体化的知识表达；

•**知识融合**：在获得新知识之后，需要对其进行整合，以消除矛盾和歧义，比如某些实体可能有多种表达，某个特定称谓也许对应于多个不同的实体等；

•**知识加工**：对于经过融合的新知识，需要经过质量评估之后（部分需要人工参与甄别），才能将合格的部分加入到知识库中，以确保知识库的质量。

知识图谱的构建：信息抽取

- 信息抽取（information extraction）是一种自动化地从半结构化和无结构数据中抽取实体、关系以及实体属性等结构化信息的技术
- 信息抽取是知识图谱构建的第一步，关键问题是：如何从异构数据源中自动抽取信息得到候选单元
- 涉及的关键技术包括：实体抽取、关系抽取和属性抽取

信息抽取：实体抽取

- 实体抽取又称为命名实体识别（named entity recognition，NER），是指从文本数据集中自动识别出命名实体
- 实体抽取的质量（准确率和召回率）对后续的知识获取效率和质量影响极大，因此是信息抽取中最为基础和关键的部分

信息抽取：关系抽取

- 文本语料经过实体抽取，得到的是离散的命名实体，为了得到语义信息，还需要从相关的语料中提取出实体之间的关联关系，通过关联关系将实体（概念）联系起来，形成网状的知识结构，研究关系抽取技术的目的，就是解决如何从文本语料中抽取实体间的关系这一基本问题

信息抽取：属性抽取

- 属性抽取(Attribute Extraction)的目标是从不同信息源中采集特定实体的属性信息
- 例如针对某个公众人物，可以从网络公开信息中得到其昵称、生日、国籍、教育背景等信息

知识图谱的构建：知识融合

- 知识融合包括：实体链接和知识合并
  - **实体链接**（entity linking）：是指对于从文本中抽取得到的实体对象，将其链接到知识库中对应的正确实体对象的操作
  - 其基本思想是首先根据给定的实体指称项，从知识库中选出一组候选实体对象，然后通过**相似度计算**将指称项链接到正确的实体对象

- 实体链接的流程：

1. 从文本中通过实体抽取得到实体指称项
2. 进行**实体消歧**和**共指消解**，判断知识库中的同名实体与之是否代表不同的含义，以及是否有其他实体与之表示相同的含义
3. 在确认知识库中对应的正确实体对象之后，将该实体指称项链接到知识库中对应实体
4. **实体消歧**：专门用于解决同名实体产生歧义问题的技术，通过实体消歧，就可以根据当前的语境，准确建立实体链接，实体消歧主要采用聚类法
5. **共指消解**：**主要用于解决多个指称对应同一实体对象的问题。**在一次会话中，多个指称可能指向的是同一实体对象。利用共指消解技术，可以将这些指称项关联到正确的实体对象

❑ 知识合并

- 构建知识图谱时，可从第三方知识库或结构化数据获取输入
- 将外部知识库融合到本地知识库需要处理两个层面的问题：
  - 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
  - 通过模式层的融合，将新得到的本体融入已有的本体库中
- 然后是合并关系数据库，在知识图谱构建过程中，一个重要的高质量知识来源是企业或者机构自己的关系数据库。为了将这些结构化的历史数据融入到知识图谱中，可以采用资源描述框架（RDF）作为数据模型，其实质就是将关系数据库的数据换成RDF的三元组数据

知识图谱的构建：知识加工

❑ 通过信息抽取，从原始语料中提取出了实体、关系与属性等知识要素；经过知识融合，消除实体指称项与实体对象之间的歧义，得到一系列基本的事实表达

❑ 然而事实本身并不等于知识。要想最终获得结构化、网络化的知识体系，还需要经历知识加工的过程

❑ 知识加工主要包括：**本体构建**、**知识推理**和**质量评估**

知识图谱的架构

知识图谱在逻辑上可分为**模式层**与**数据层**两个层次。

模式层：

- **模式层**构建在数据层之上，是知识图谱的核心，通常采用本体库来管理知识图谱的模式层
- 本体是结构化知识库的概念模板，通过本体库而形成的知识库不仅层次结构较强，并且冗余程度较小

模式层：实体-关系-实体，实体-属性-值

知识图谱管理

知识图谱的管理，主要是知识库的更新，包括**概念层的更新**和**数据层的更新**：

- 概念层的更新是指新增数据后获得了新的概念，需要自动将新的概念添加到知识库的概念层中
- 数据层的更新主要是新增或更新实体、关系、属性值，对数据层进行更新需要考虑数据源的可靠性、数据的一致性（是否存在矛盾或冗杂等问题）等可靠数据源，并选择在各数据源中出现频率高的事实和属性加入知识库

问答系统历史

基于信息检索的问答

基于关键词匹配+信息抽取，基于浅层语义分析

Text REtrieval Conference (TREC)

基于社区的问答

依赖于网民贡献，问答过程依赖于关键词检索技术

YAHOO!

基于知识库的问答

知识库
语义解析

机器阅读理解的方法

- 传统特征工程的方法
  - 文本分析
  - 问句解析
  - 匹配答案
- 神经网络的方法
  - 文档和问句的表示学习
  - 文档和问句的匹配计算
  - 深度推理机制

趋势热点：值得关注的 NLP 技术