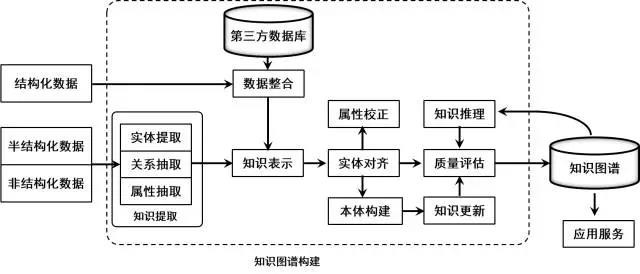
知识图谱简介

# 简介

## 知识图谱构建

知识图谱的构建架构图如下：



知识图谱在逻辑上可分为模式层与数据层两个层次，数据层主要是由一系列的事实组成，而知识将以事实为单位进行存储。如果用(实体1，关系，实体2)、(实体、属性，属性值)这样的三元组来表达事实，可选择图数据库作为存储介质，例如开源的Neo4j[7]、Twitter的FlockDB[8]、sones的GraphDB[9]等。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。知识图谱在逻辑上可分为模式层与数据层两个层次，数据层主要是由一系列的事实组成，而知识将以事实为单位进行存储。如果用(实体1，关系，实体2)、(实体、属性，属性值)这样的三元组来表达事实，可选择图数据库作为存储介质，例如开源的Neo4j[7]、Twitter的FlockDB[8]、sones的GraphDB[9]等。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。

实体: 指的是具有可区别性且独立存在的某种事物。如某一个人、某一个城市、某一种植物等、某一种商品等等。世界万物有具体事物组成，此指实体。如图1的“中国”、“美国”、“日本”等。，实体是知识图谱中的最基本元素，不同的实体间存在不同的关系。

语义类（概念）：具有同种特性的实体构成的集合，如国家、民族、书籍、电脑等。 概念主要指集合、类别、对象类型、事物的种类，例如人物、地理等。

内容: 通常作为实体和语义类的名字、描述、解释等，可以由文本、图像、音视频等来表达。

属性(值): 从一个实体指向它的属性值。不同的属性类型对应于不同类型属性的边。属性值主要指对象指定属性的值。如图1所示的“面积”、“人口”、“首都”是几种不同的属性。属性值主要指对象指定属性的值，例如960万平方公里等。

关系: 形式化为一个函数，它把kk个点映射到一个布尔值。在知识图谱上，关系则是一个把kk个图节点(实体、语义类、属性值)映射到布尔值的函数。

### 自顶向下

自顶向下指的是先为知识图谱定义好本体与数据模式，再将实体加入到知识库。该构建方式需要利用一些现有的结构化知识库作为其基础知识库，例如Freebase项目就是采用这种方式，它的绝大部分数据是从维基百科中得到的。

### 自底向上

自底向上指的是从一些开放链接数据中提取出实体，选择其中置信度较高的加入到知识库，再构建顶层的本体模式。目前，大多数知识图谱都采用自底向上的方式进行构建，其中最典型就是Google的Knowledge Vault[11]和微软的Satori知识库。现在也符合互联网数据内容知识产生的特点。

# 关键技术

## 知识提取

### 基于百科或垂直站点提取

### 基于规则与词典的实体提取方法

### 基于统计机器学习的实体抽取方法

### 面向开放域的实体抽取方法

## 语义类抽取

第一步：并列相似度计算

第二步：上下位关系提取

第三步：语义类生成

## 属性和属性值提取

# 应用

## 百度知识图谱