

基于多模态知识图谱的 系统设计

针对多模态知识图谱构建与存储方面



杨卉帆

2022-4-30

**目录**

[1 实践目的 2](#_Toc102223658)

多模态知识图谱的构建



# 知识图谱简介

知识图谱，是结构化的语义知识库，用于迅速描述物理世界中的概念及其相互关系，通过知识图谱能够将Web上的信息、数据以及链接关系聚集为知识，使信息资源更易于计算、理解以及评价，并能实现知识的快速响应和推理。

## 广泛应用于各领域

当下知识图谱已在工业领域得到了广泛应用，如搜索领域的Google搜索、百度搜索，社交领域的领英经济图谱，企业信息领域的天眼查企业图谱，电商领域的淘宝商品图谱，O2O领域的美团知识大脑，医疗领域的丁香园知识图谱，以及工业制造业知识图谱等。

## 构建技术分类

在知识图谱技术发展初期，很多企业和科研机构会采用自顶向下的方式构建基础知识库，如 Freebase。随着自动知识抽取与加工技术的不断成熟，当前的知识图谱大多采用自底向上的方式构建，如 Google 的 Knowledge Vault 和微软的 Satori 知识库。

知识图谱的构建技术主要有自顶向下和自底向上两种。

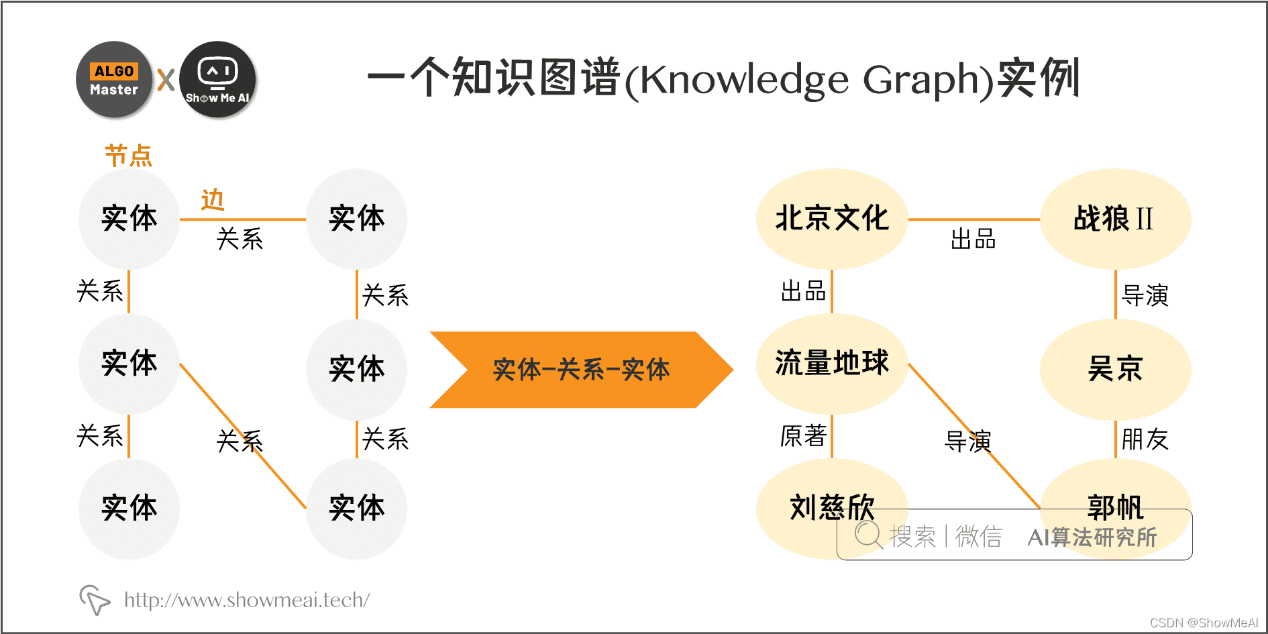
1. **自顶向下构建**：借助百科类网站等结构化数据源，从高质量数据中提取本体和模式信息，加入到知识库里。
2. **自底向上构建**：借助一定的技术手段，从公开采集的数据中提取出资源模式，选择其中置信度较高的信息，加入到知识库中。



## “实体-关系-实体”三元组

下图是典型的知识图谱样例示意图。可以看到，“图谱”中有很多节点，如果两个节点之间存在关系，他们就会被一条边连接在一起，这个节点我们称为**实体（Entity）**，节点之间的这条边，我们称为**关系（Relationship）**。

知识图谱的基本单位，就是“**实体(Entity)-关系(Relationship)-实体(Entity)**” 构成的三元组，这也是知识图谱的核心。



# 数据类型和存储方式

## 三类数据

知识图谱的原始数据类型一般来说有三类（也是互联网上的三类原始数据）：

* **结构化数据（Structured Data）**，如：关系数据库、链接数据
* **半结构化数据（Semi-Structured Data）**，如：XML、JSON、百科
* **非结构化数据（Unstructured Data）**，如：图片、音频、视频



典型的半结构化数据样例如下：



## 存储方式：RDF（资源描述框架），图数据库

如何存储上面这三类数据类型呢？一般有两种选择：

1. 可以通过RDF（资源描述框架）这样的规范存储格式来进行存储，比较常用的有Jena等。

<RDF>

<Description about="https://www.w3.org/RDF/">

<author>HanXinzi</author>

<homepage> http://www.showmeai.tech </homepage>

</Description>

</RDF>

1. 另一种方法是使用图数据库来进行存储，常用的有Neo4j等。



**截止目前为止，看起来知识图谱主要是一堆三元组，那用关系数据库来存储可以吗？**

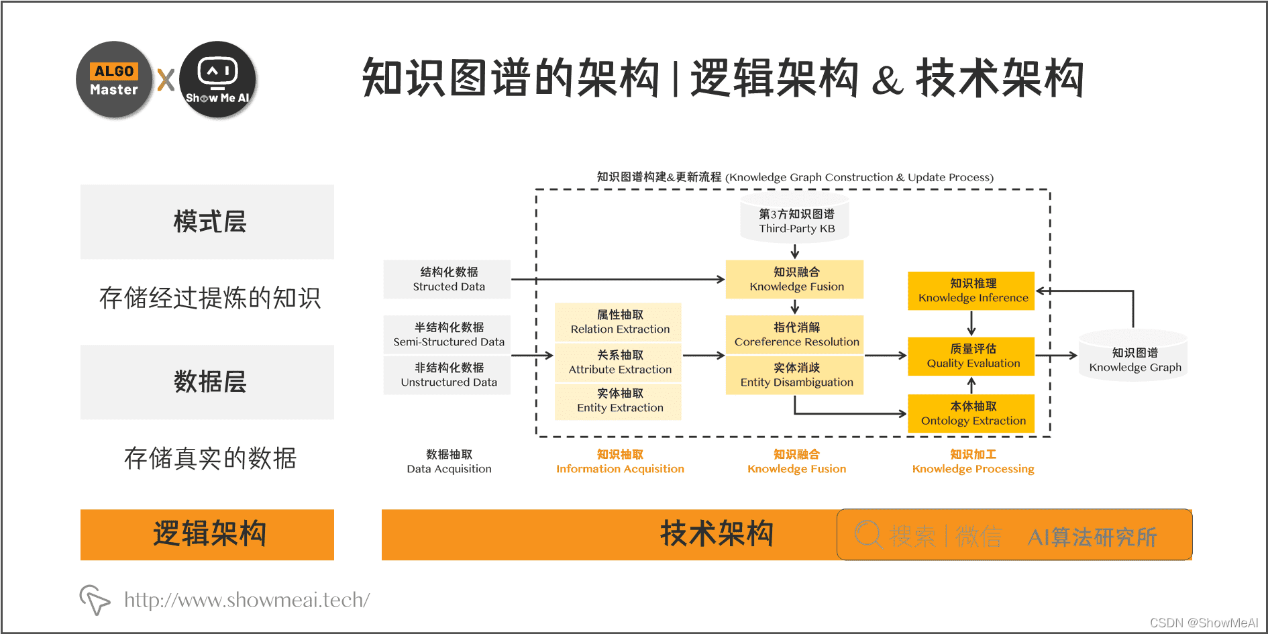
对，从技术上来说，用关系数据库来存储知识图谱（尤其是简单结构的知识图谱），是完全没问题的。但一旦知识图谱变复杂，用传统的「关系数据存储」，查询效率会显著低于「图数据库」。在一些涉及到2,3度的关联查询场景，图数据库能把查询效率提升几千倍甚至几百万倍。

而且基于图的存储在设计上会非常灵活，一般只需要局部的改动即可。当你的场景数据规模较大的时候，建议直接用图数据库来进行存储。

# 知识图谱的架构

知识图谱的架构主要可以被分为：

* **逻辑架构**
* **技术架构**



## 逻辑架构

在逻辑上，我们通常将知识图谱划分为两个层次：数据层和模式层。

* **模式层**：在数据层之上，是知识图谱的核心，存储经过提炼的知识，通常通过本体库来管理这一层（本体库可以理解为面向对象里的“类”这样一个概念，本体库就储存着知识图谱的类）。
* **数据层**：存储真实的数据。

**可以看看这个例子：**  
 模式层：实体-关系-实体，实体-属性-性值  
 数据层：吴京-妻子-谢楠，吴京-导演-战狼Ⅱ

## 技术架构

知识图谱的整体架构如图所示，其中虚线框内的部分为知识图谱的构建过程，同时也是知识图谱更新的过程。下面让我们顺着这张图来理一下思路。

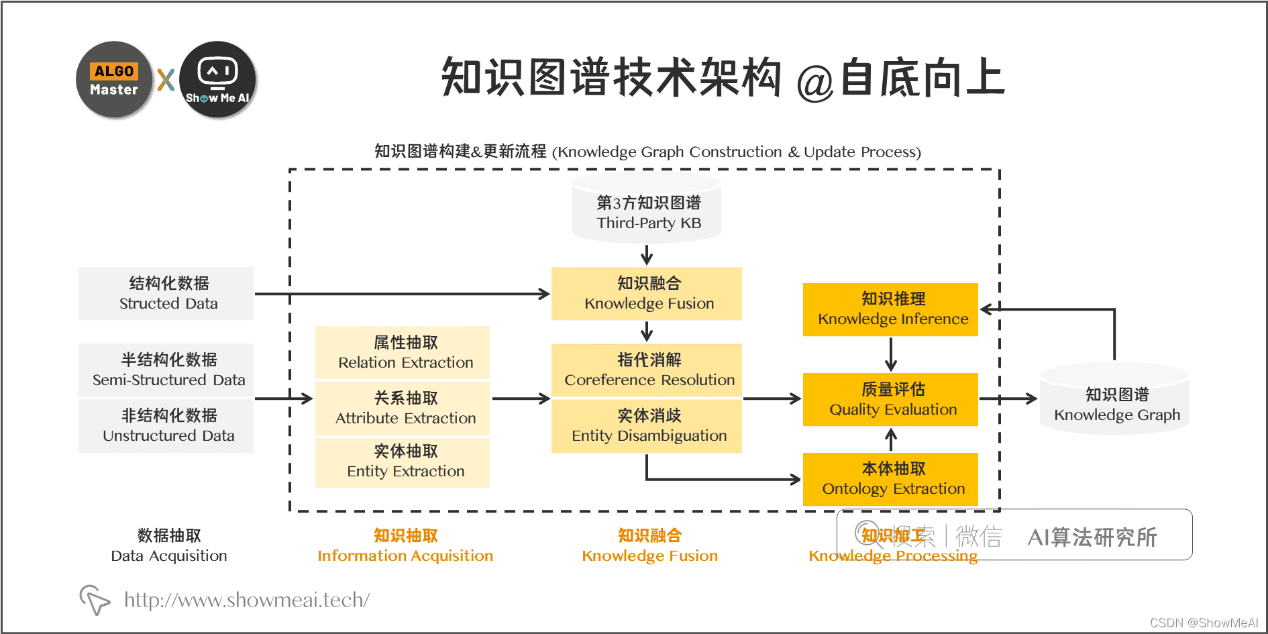
* 首先，我们有一大堆的数据，这些数据可能是结构化的、非结构化的以及半结构化的；
* 然后，我们基于这些数据来构建知识图谱，这一步主要是通过一系列自动化或半自动化的技术手段，来从原始数据中提取出知识要素，即一堆实体关系，并将其存入我们的知识库的模式层和数据层。

# 构建技术

前面的内容说到了，知识图谱有自顶向下和自底向上两种构建方式，这里提到的构建技术主要是**自底向上**的构建技术。

如前所述，构建知识图谱是一个迭代更新的过程，根据知识获取的逻辑，每一轮迭代包含三个阶段：

* **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（需要实体消歧），某个特定称谓也许对应于多个不同的实体等（需要指代消解）。
* **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。

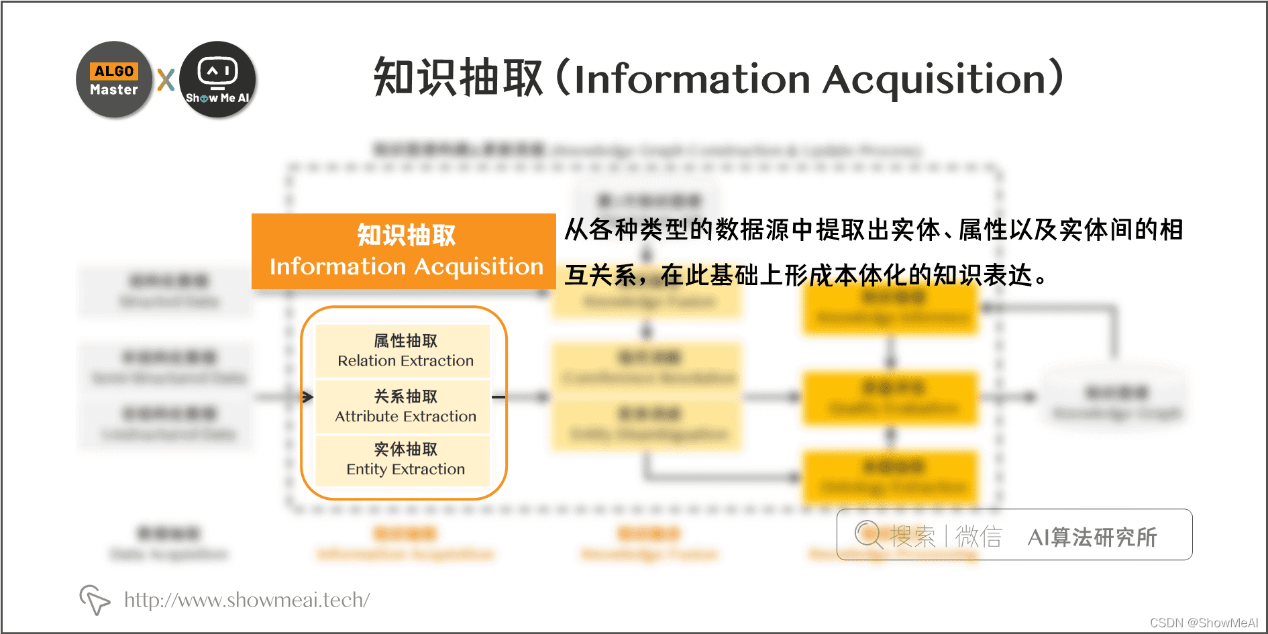


下面我们依次对每一个步骤进行介绍。

## 知识抽取

知识抽取（information extraction）是知识图谱构建的第1步，其中的关键问题是：**如何从异构数据源中自动抽取信息得到候选指示单元？**

信息抽取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。涉及的关键技术包括：**实体抽取**、**关系抽取**和**属性抽取**。



### 实体抽取

**实体抽取，也称为命名实体识别（named entity recognition，NER），是指从文本数据集中自动识别出命名实体。**

下图中，通过实体抽取我们可以从其中抽取出四个实体：“非洲”、“中国海军”、“冷锋”、“战狼”。



**研究历史**：从面向单一领域进行实体抽取，逐步跨步到面向开放域（Open Domain）的实体抽取。

### 关系抽取

文本语料经过实体抽取之后，得到的是一系列离散的命名实体。**为了得到语义信息，还需要从相关语料中提取出实体之间的关联关系，通过关系将实体联系起来，才能够形成网状的知识结构**。这就是关系抽取需要做的事，如下图所示。

# 参考文献

* 大厂技术实现 | 详解知识图谱的构建全流程 @自然语言处理系列  
  <https://www.1024sou.com/article/677171.html>