

基于多模态知识图谱的 系统设计

数据采集-多模态知识图谱构建与存储-应用服务



杨卉帆

2022-5

目录

[架构图模板 -基于多模态知识图谱的系统设计 3](#_Toc102490170)

[第一篇：知识图谱的构建与存储 4](#_Toc102490171)

[1 知识图谱简介 4](#_Toc102490172)

[1.1 广泛应用于各领域 4](#_Toc102490173)

[1.2 构建技术分类 4](#_Toc102490174)

[1.3 “实体-关系-实体”三元组 5](#_Toc102490175)

[2 数据类型和存储方式 6](#_Toc102490176)

[2.1 三类数据 6](#_Toc102490177)

[2.2 存储方式：RDF（资源描述框架），图数据库 7](#_Toc102490178)

[3 知识图谱的架构 8](#_Toc102490179)

[3.1 逻辑架构 8](#_Toc102490180)

[3.2 技术架构 8](#_Toc102490181)

[4 构建技术 9](#_Toc102490182)

[4.1 知识抽取 10](#_Toc102490183)

[4.1.1 实体抽取 10](#_Toc102490184)

[4.1.2 关系抽取 11](#_Toc102490185)

[4.1.3 属性抽取 11](#_Toc102490186)

[4.2 知识融合 12](#_Toc102490187)

[4.2.1 实体链接 13](#_Toc102490188)

[4.2.2 知识融合 14](#_Toc102490189)

[4.3 知识加工 15](#_Toc102490190)

[4.3.1 本体抽取 16](#_Toc102490191)

[4.3.2 知识推理 17](#_Toc102490192)

[4.3.3 质量评估 18](#_Toc102490193)

[4.4 知识更新 18](#_Toc102490194)

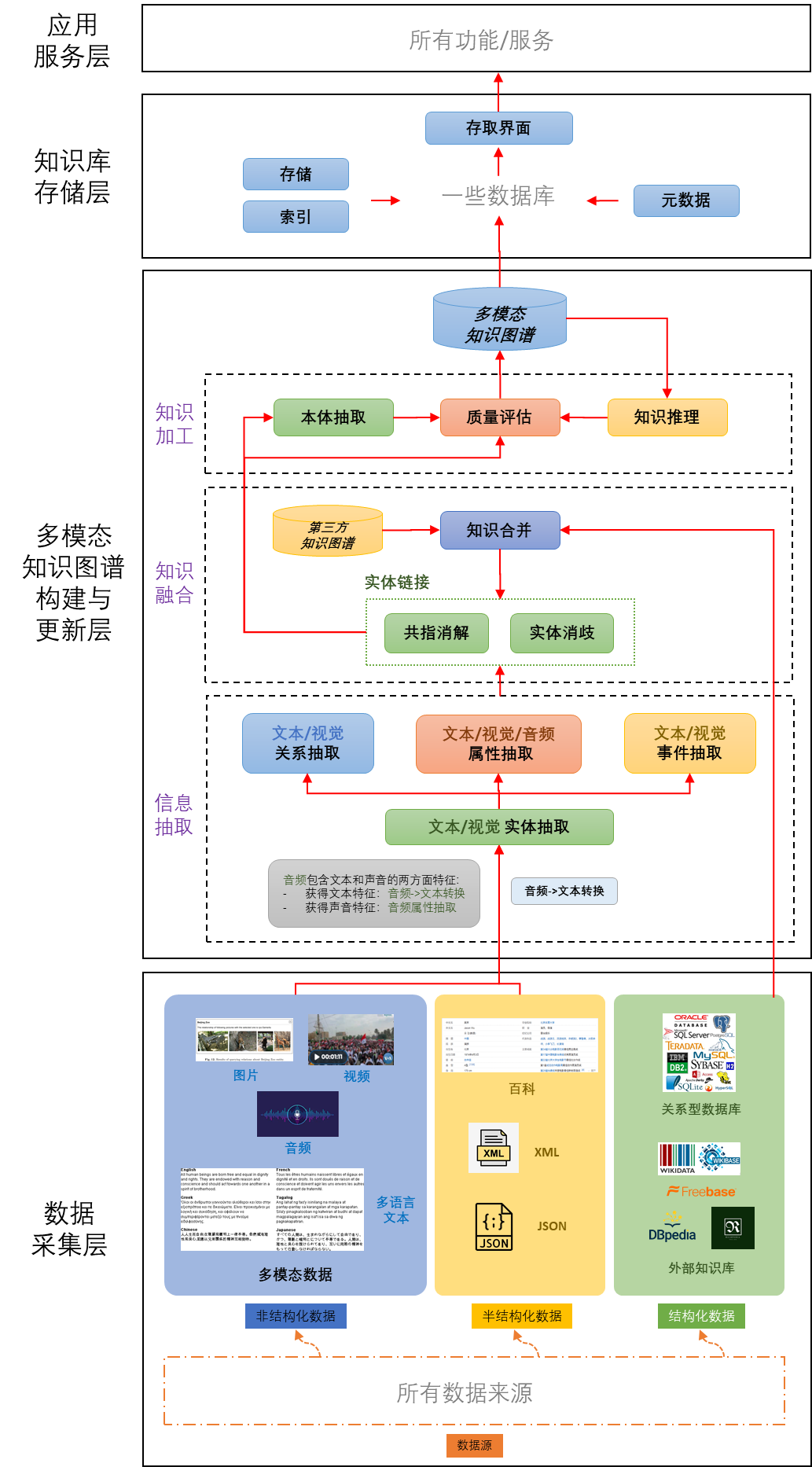
[第二篇：『多模态』知识图谱构建 20](#_Toc102490195)

[1 GAIA: A Fine-grained Multimedia Knowledge Extraction System 20](#_Toc102490196)

[2 Information from audio 21](#_Toc102490197)

[第三篇：架构图模板 -基于多模态知识图谱的系统设计 23](#_Toc102490198)

架构图模板  
-基于多模态知识图谱的系统设计



第一篇：知识图谱的构建与存储

<http://www.showmeai.tech/article-detail/knowledge-graph>



# 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。

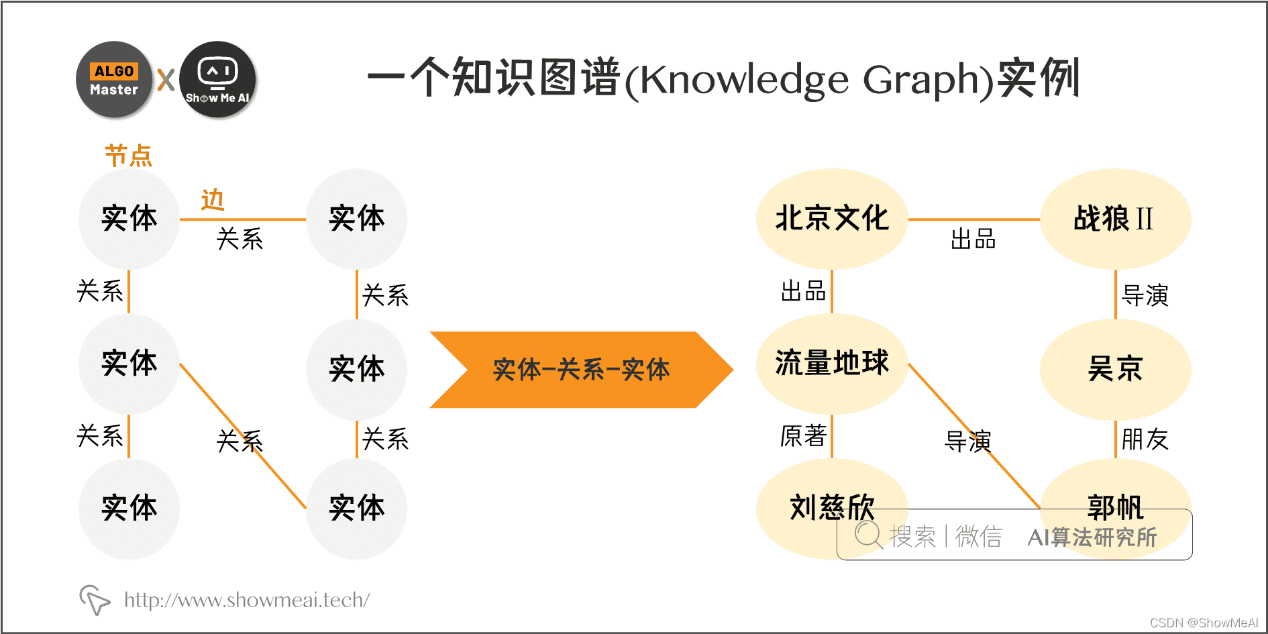


\*本体：在[计算机科学](https://zh.wikipedia.org/wiki/%E8%AE%A1%E7%AE%97%E6%9C%BA%E7%A7%91%E5%AD%A6)和[信息科学](https://zh.wikipedia.org/wiki/%E4%BF%A1%E6%81%AF%E7%A7%91%E5%AD%A6)中，**本体**是指对概念、数据和实体之间的类别、属性和关系的表示、命名和定义，这些概念、数据和[实体](https://zh.wikipedia.org/wiki/%E5%AE%9E%E4%BD%93)构成了一个、大量或所有的[论域](https://zh.wikipedia.org/wiki/%E8%AE%BA%E5%9F%9F)[[1]](https://zh.wikipedia.org/wiki/%E6%9C%AC%E4%BD%93_(%E4%BF%A1%E6%81%AF%E7%A7%91%E5%AD%A6)#cite_note-1)。

## “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。

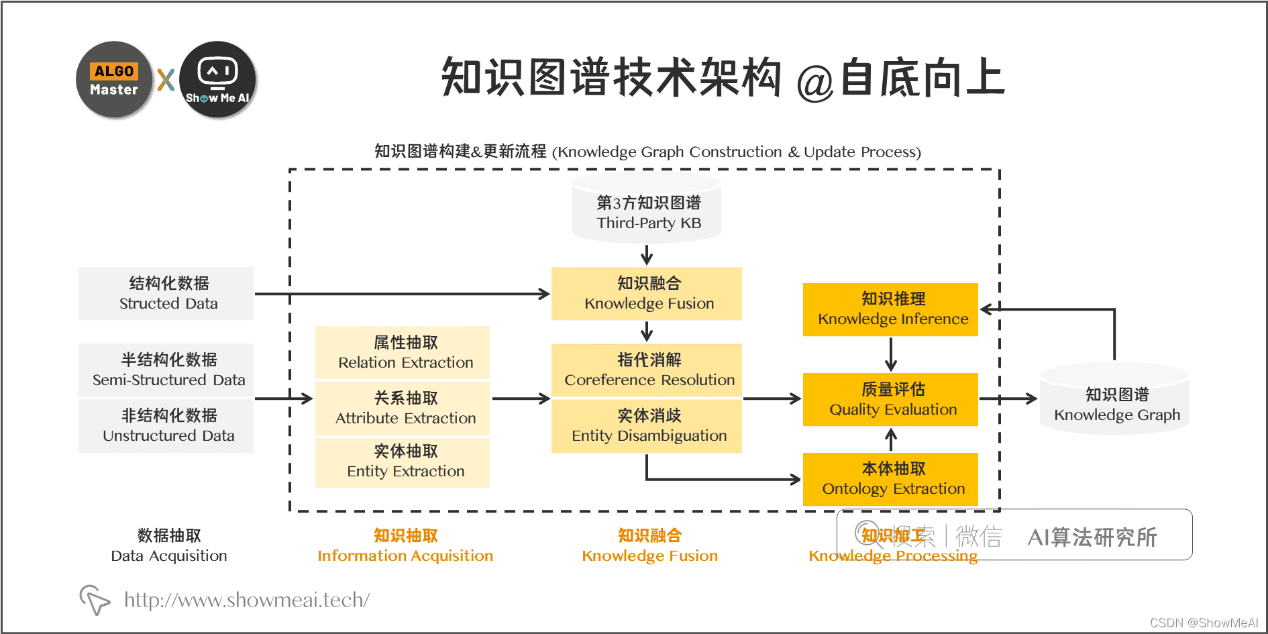
* 首先，我们有一大堆的数据，这些数据可能是结构化的、非结构化的以及半结构化的；
* 然后，我们基于这些数据来构建知识图谱，这一步主要是通过一系列自动化或半自动化的技术手段，来从原始数据中提取出知识要素，即一堆实体、关系和属性，并将其存入我们的知识库的模式层和数据层。

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

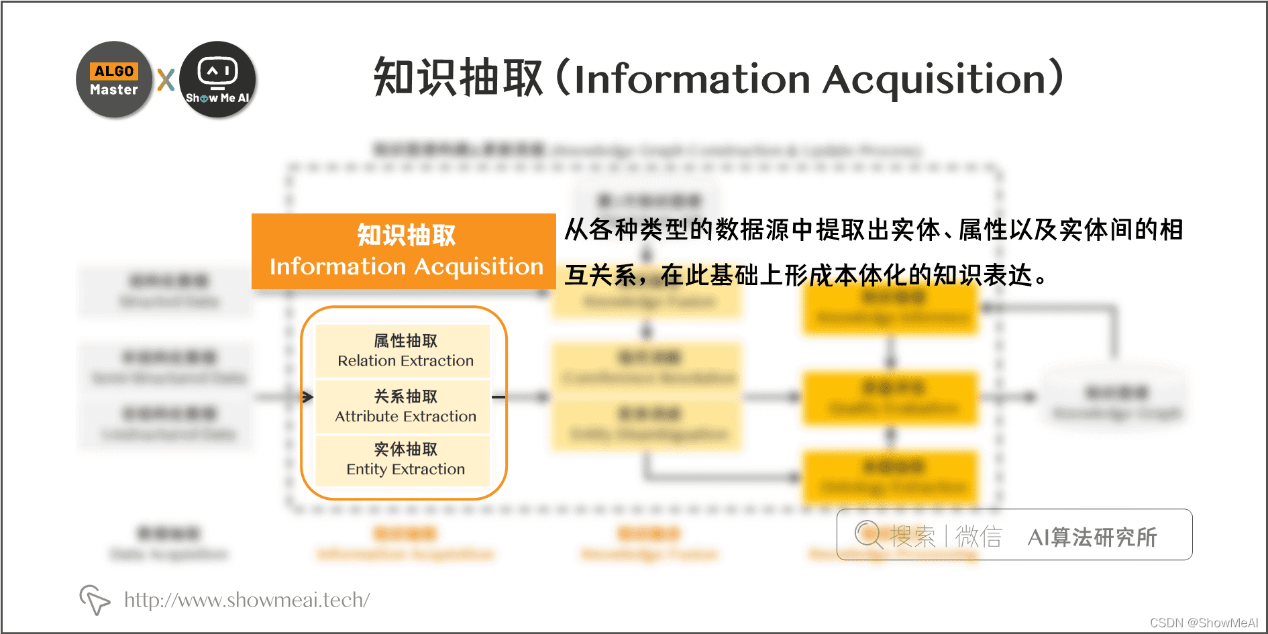


下面我们依次对每一个步骤进行介绍。

## 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。**

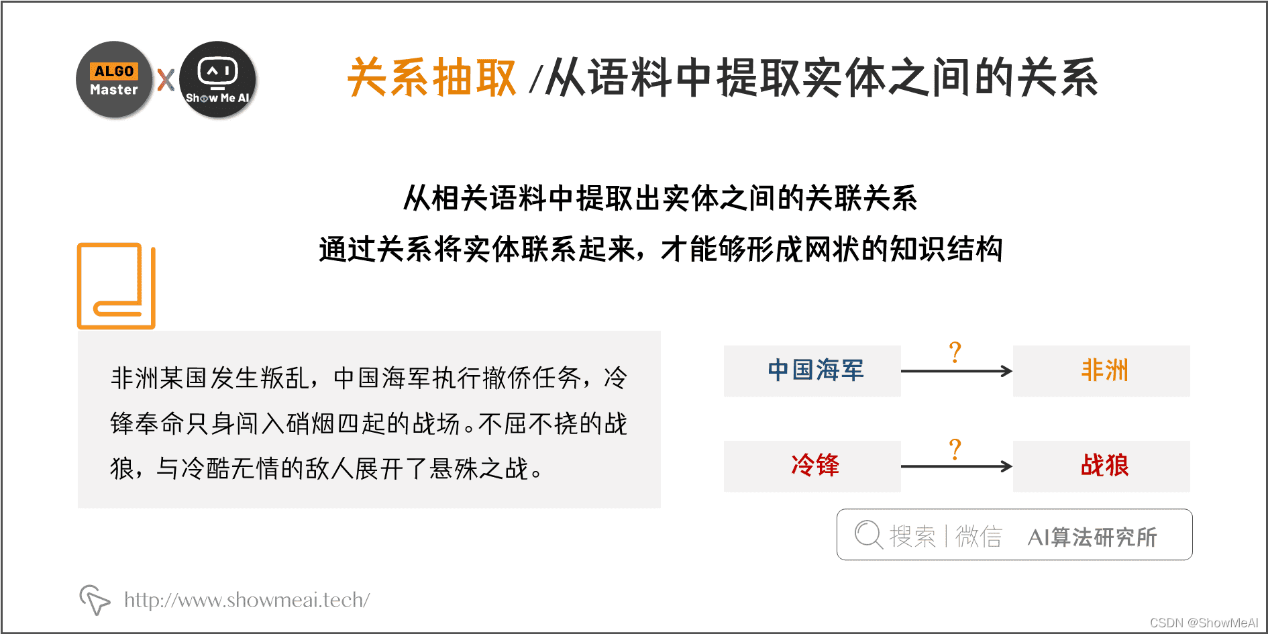
下图中，通过实体抽取我们可以从其中抽取出四个实体：“非洲”、“中国海军”、“冷锋”、“战狼”。



**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。

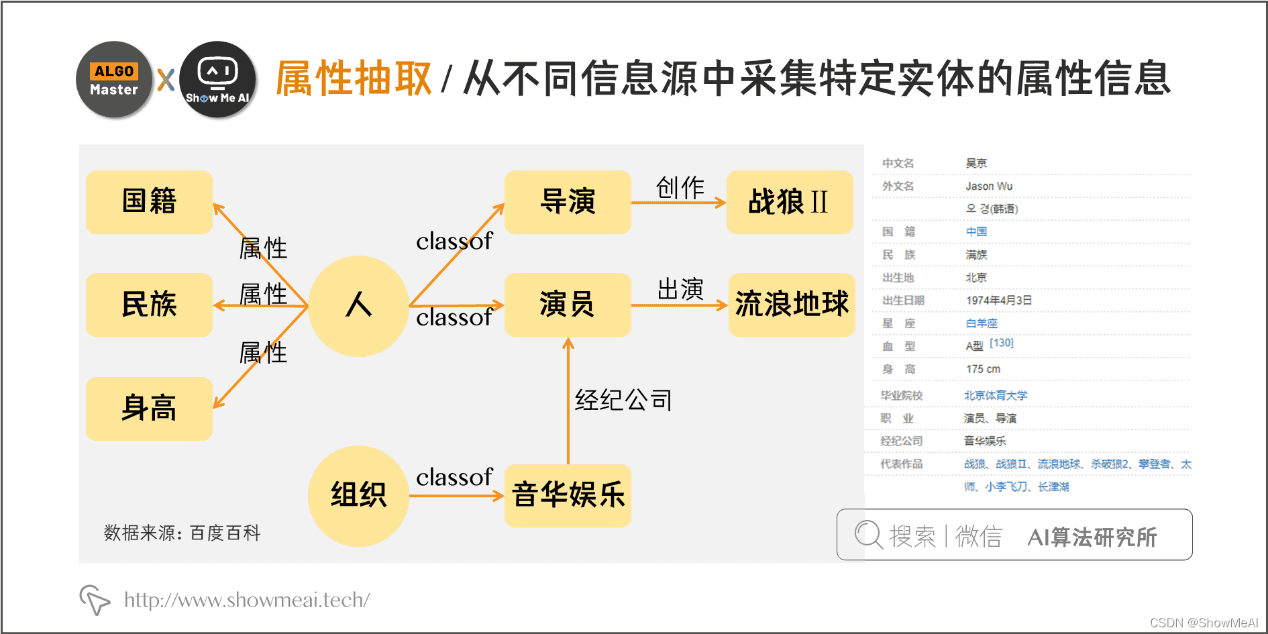


**研究历史**：

人工构造语法和语义规则（模式匹配）。  
统计机器学习方法。  
基于特征向量或核函数的有监督学习方法。  
研究重点转向半监督和无监督。  
开始研究面向开放域的信息抽取方法。  
将面向开放域的信息抽取方法和面向封闭领域的传统方法结合。

### 属性抽取

属性抽取的目标是从不同信息源中采集特定实体的属性信息，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。



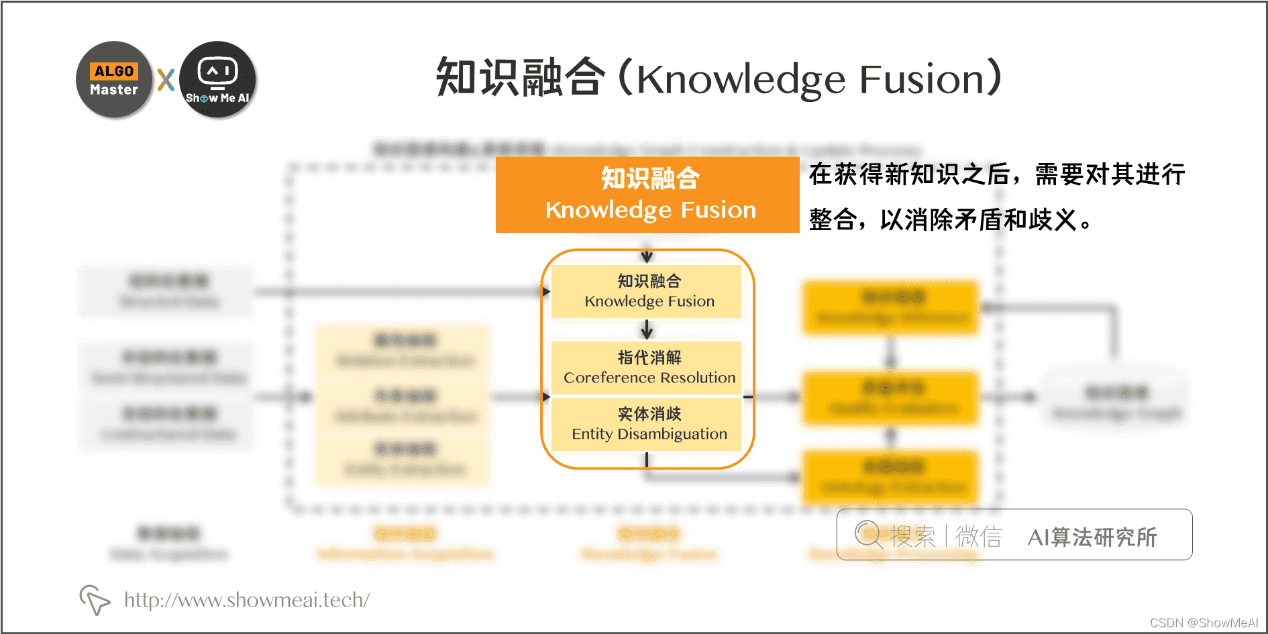
**研究历史**：  
- 将实体的属性视作实体与属性值之间的一种名词性关系，将属性抽取任务转化为关系抽取任务。  
- 基于规则和启发式算法，抽取结构化数据。  
- 基于百科类网站的半结构化数据，通过自动抽取生成训练语料，用于训练实体属性标注模型，然后将其应用于对非结构化数据的实体属性抽取。  
- 采用数据挖掘的方法直接从文本中挖掘实体属性和属性值之间的关系模式，据此实现对属性名和属性值在文本中的定位。

## 知识融合

通过信息抽取，我们就从原始的非结构化和半结构化数据中获取到了实体、关系以及实体的属性信息。如果我们将接下来的过程比喻成拼图的话，那么这些信息就是拼图碎片，散乱无章甚至还有从其他拼图里跑来的碎片、本身就是用来干扰我们拼图的错误碎片。

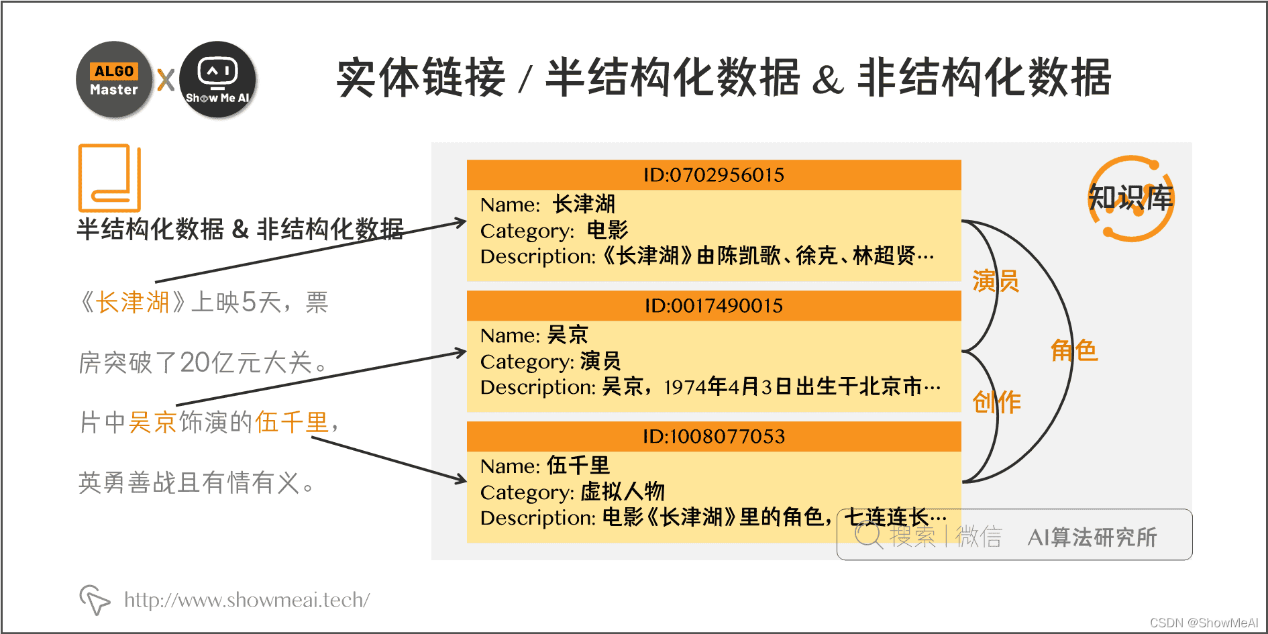
也就是说，拼图碎片（信息）之间的关系是扁平化的，缺乏层次性和逻辑性；拼图（知识）中还存在大量冗杂和错误的拼图碎片（信息）。那么如何解决这一问题，就是在知识融合这一步里我们需要做的了。

**知识融合包括2部分内容：实体链接、知识合并。**



### 实体链接

**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。



**研究历史**：  
- 仅关注如何将从文本中抽取到的实体链接到知识库中，忽视了位于同一文档的实体间存在的语义联系；  
- 开始关注利用实体的共现关系，同时将多个实体链接到知识库中。即集成实体链接（collective entity linking）。

**实体链接的流程：**

* 从文本中通过实体抽取得到实体指称项。
* 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知识库中是否存在其他命名实体与之表示相同的含义。
* 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。
* **实体消歧**：是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。其实也可以看做基于上下文的分类问题，类似于词性消歧和词义消歧。
* **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（合并）到正确的实体对象，由于该问题在信息检索和自然语言处理等领域具有特殊的重要性，吸引了大量的研究努力。共指消解还有一些其他的名字，比如对象对齐、实体匹配和实体同义。

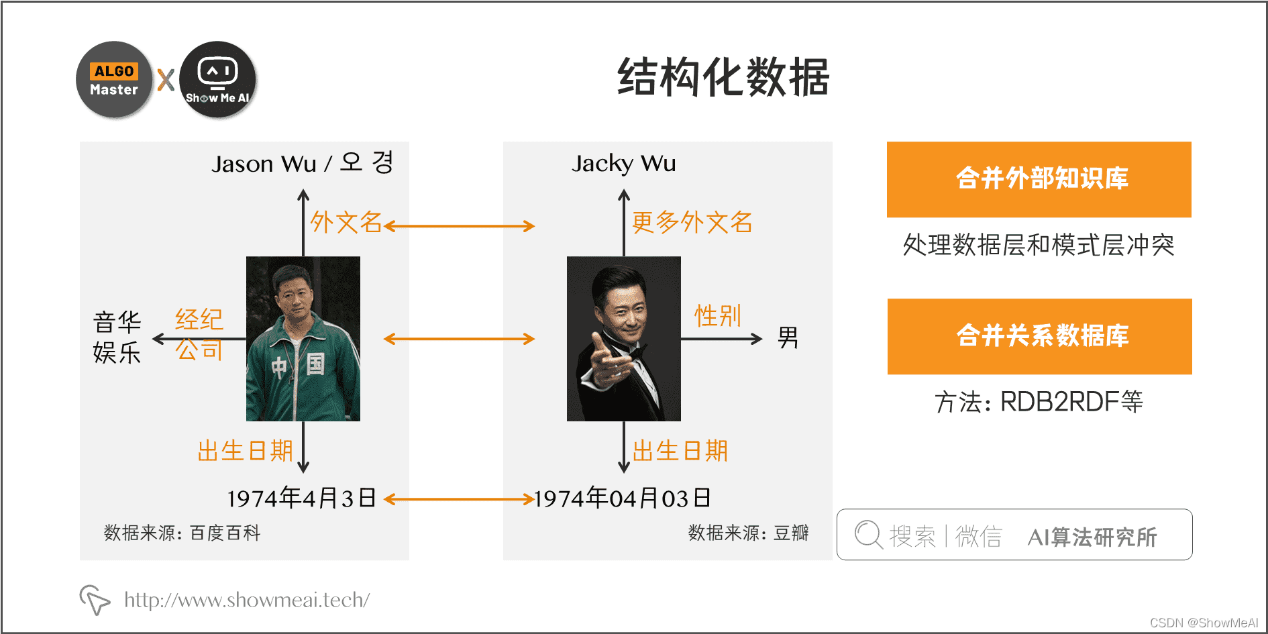
### 知识融合

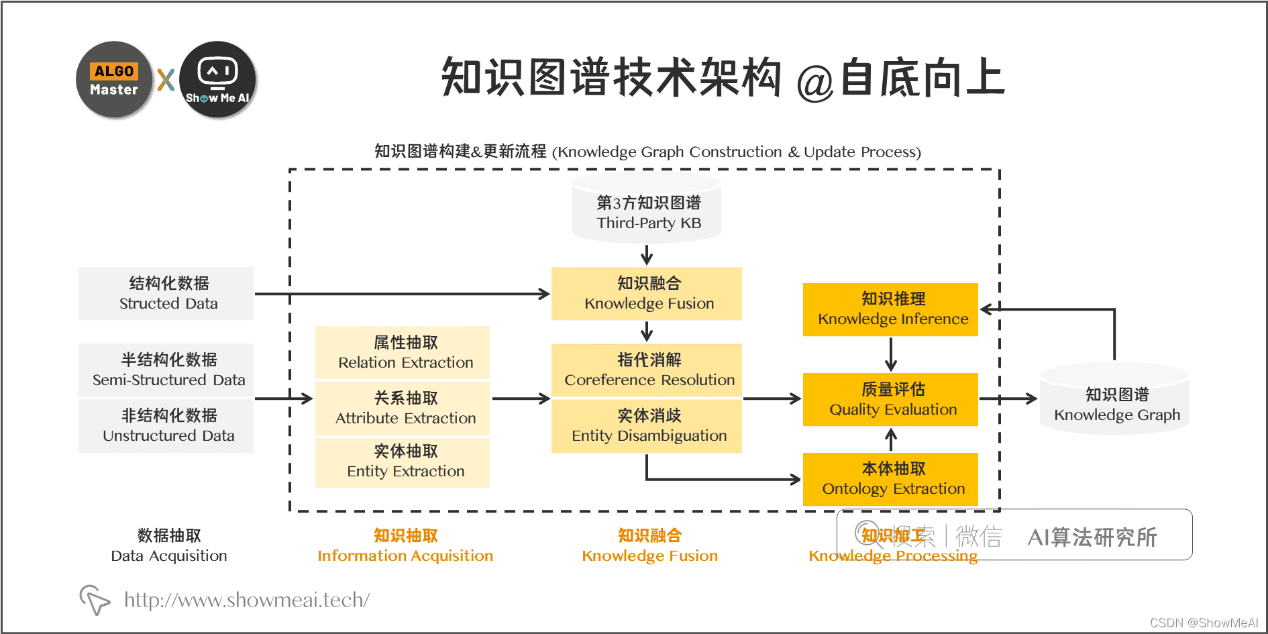
在前面的实体链接中，我们已经将实体链接到知识库中对应的正确实体对象那里去了，但需要注意的是，实体链接链接的是我们从半结构化数据和非结构化数据那里通过信息抽取提取出来的数据。

那么除了半结构化数据和非结构化数据以外，我们还有个更方便的数据来源———**结构化数据**，如外部知识库和关系数据库。对于这部分结构化数据的处理，就是我们知识融合的内容啦。

一般来说知识融合主要分为两种：

1. **合并外部知识库**，主要处理数据层和模式层的冲突；
2. **合并关系数据库**，有RDB2RDF等方法。

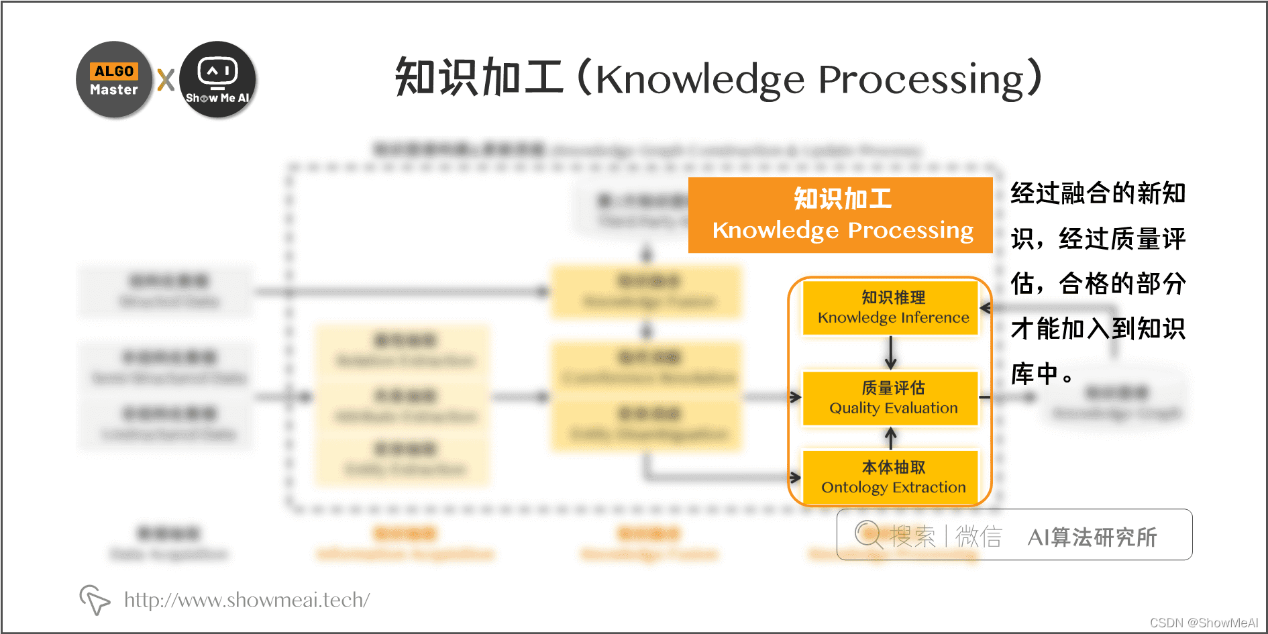




## 知识加工

经过刚才那一系列步骤，我们终于走到了知识加工这一步了！在前面，我们已经通过信息抽取，从原始语料中提取出了实体、关系与属性等知识要素，并且经过知识融合，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3方面内容：本体抽取、知识推理和质量评估。**



### 本体抽取

**本体（ontology）**是指工人的概念集合、概念框架，如“人”、“事”、“物”等。**本体可以采用人工编辑的方式手动构建（借助本体编辑软件），也可以以数据驱动的自动化方式构建本体。**因为人工方式工作量巨大，且很难找到符合要求的专家，因此当前主流的全局本体库产品，都是从一些面向特定领域的现有本体库出发，采用自动构建技术逐步扩展得到的。

**自动化本体构建过程包含三个阶段：** 实体并列关系相似度计算 → 实体上下位关系抽取 → 本体的生成。



* 第一步『实体并列关系相似度计算』  
  如图所示，当知识图谱刚得到“战狼Ⅱ”、“流浪地球”、“北京文化”这三个实体的时候，可能会认为它们三个之间并没有什么差别。但当它去计算三个实体之间的相似度后，就会发现，“战狼Ⅱ”和“流浪地球”之间可能更相似，与“北京文化”差别更大一些。
* 第二步『**实体上下位关系抽取**』  
  第一步下来，**知识图谱实际上还是没有一个上下层的概念**。它还是不知道，“流浪地球”和“北京文化”不隶属于一个类型，无法比较。因此第二步『**实体上下位关系抽取**』需要去完成这样的工作，从而生成第三步的本体。
* 第三步『本体的生成』

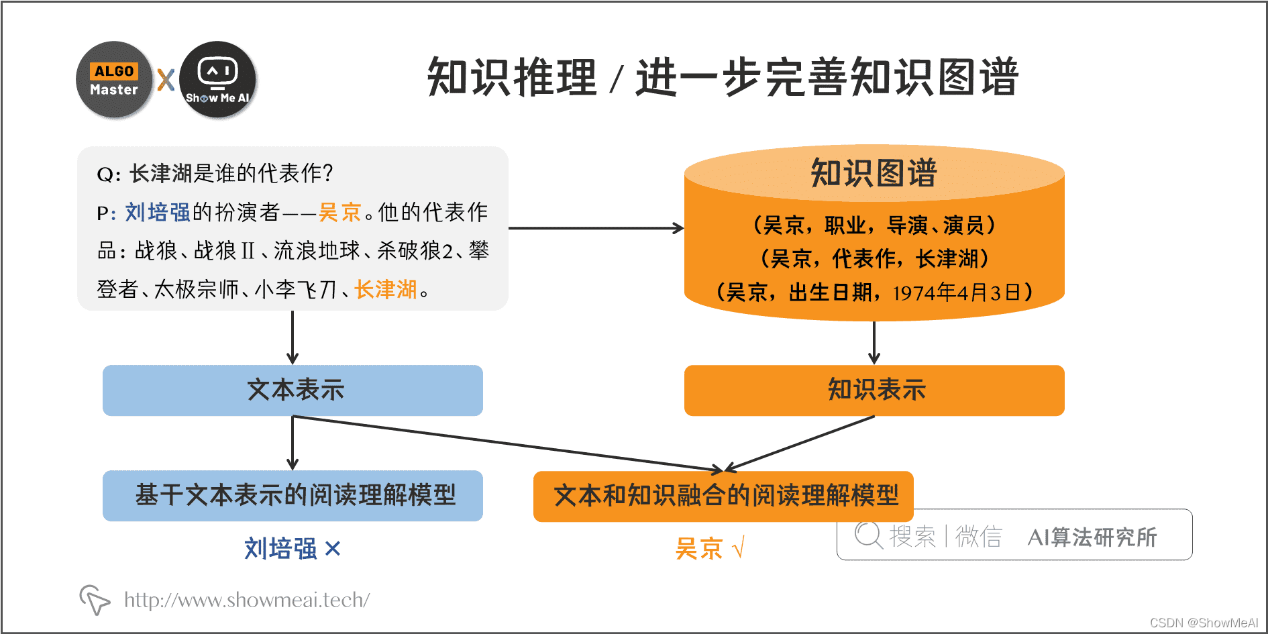
当三步结束后，这个知识图谱可能就会明白，“战狼2和流浪地球，是电影这个实体下的细分实体。它们和北京文化这家公司并不是一类”。

### 知识推理

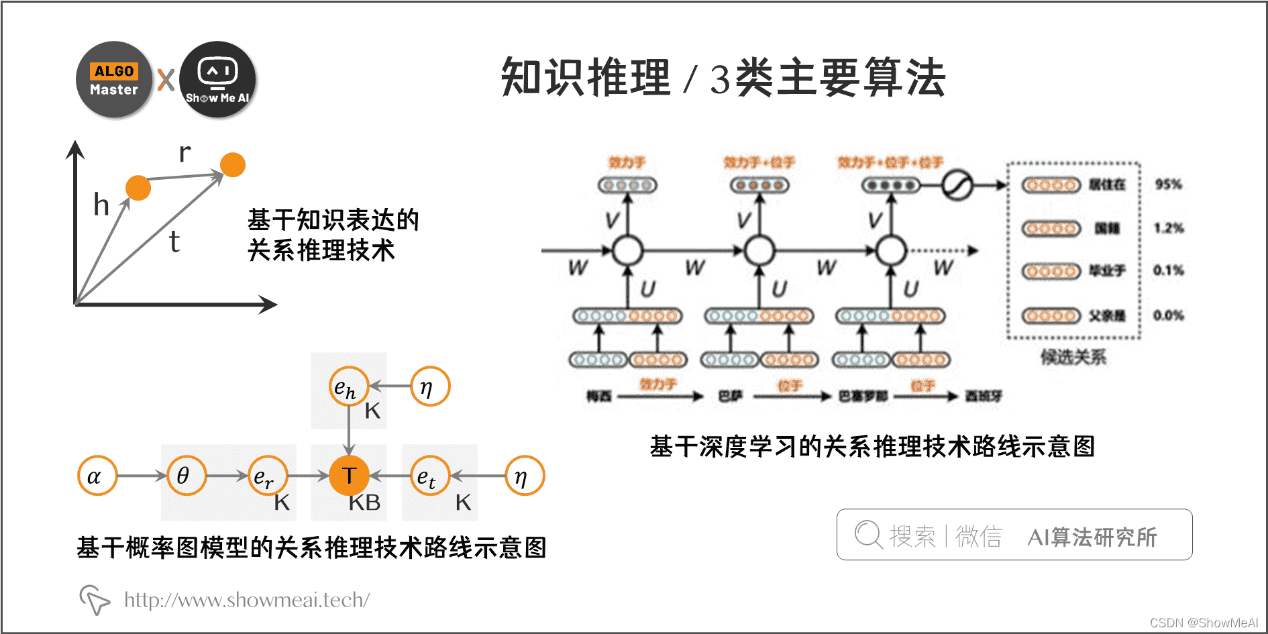
在我们完成了本体构建这一步之后，一个知识图谱的雏形便已经搭建好了。但可能在这个时候，知识图谱之间大多数关系都是残缺的，缺失值非常严重，那么这个时候，我们就可以使用知识推理技术，去完成进一步的知识发现。

当然知识推理的对象也并不局限于实体间的关系，也可以是实体的属性值，本体的概念层次关系等。

* **推理实体间的关系**
* **推理实体的属性值**：已知某实体的生日属性，可以通过推理得到该实体的年龄属性；
* **推理本体的概念层次关系**：已知(老虎，科，猫科)和（猫科，目，食肉目）可以推出（老虎，目，食肉目）



这一块的算法主要可以分为3大类：**基于知识表达的关系推理技术；基于概率图模型的关系推理技术路线示意图；基于深度学习的关系推理技术路线示意图。**



### 质量评估

质量评估也是知识库构建技术的重要组成部分，这一部分存在的意义在于：可以对知识的可信度进行量化，通过舍弃置信度较低的知识来保障知识库的质量。

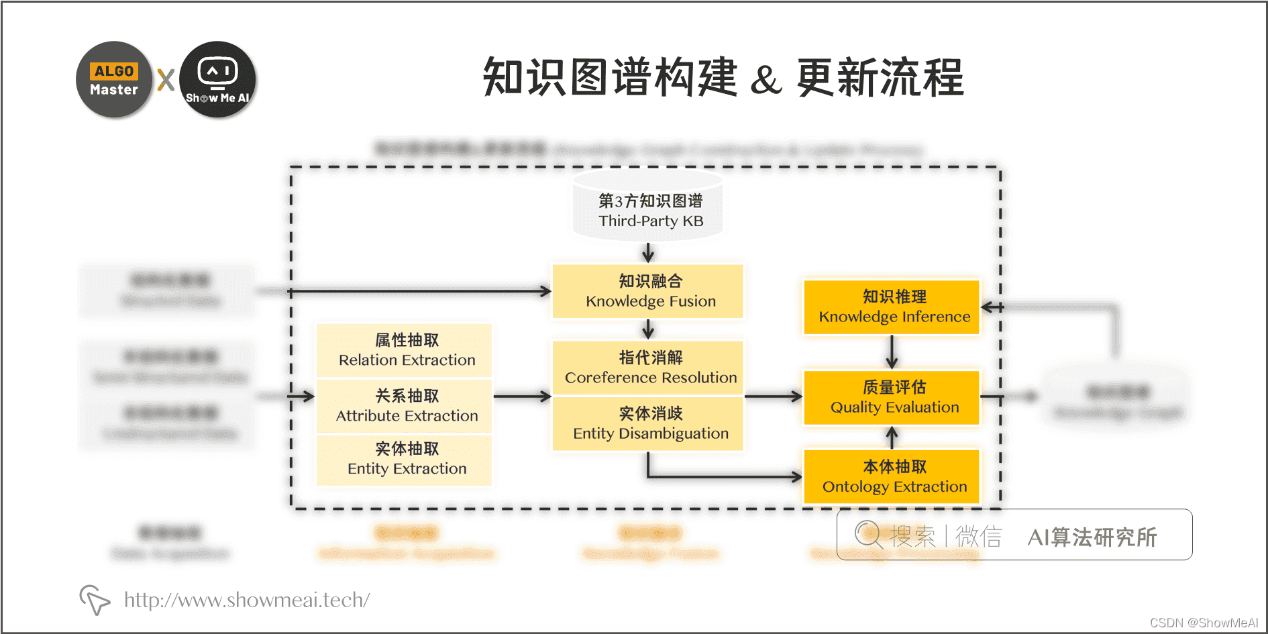
## 知识更新

从逻辑上看，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**：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。

知识图谱的内容更新有两种方式：

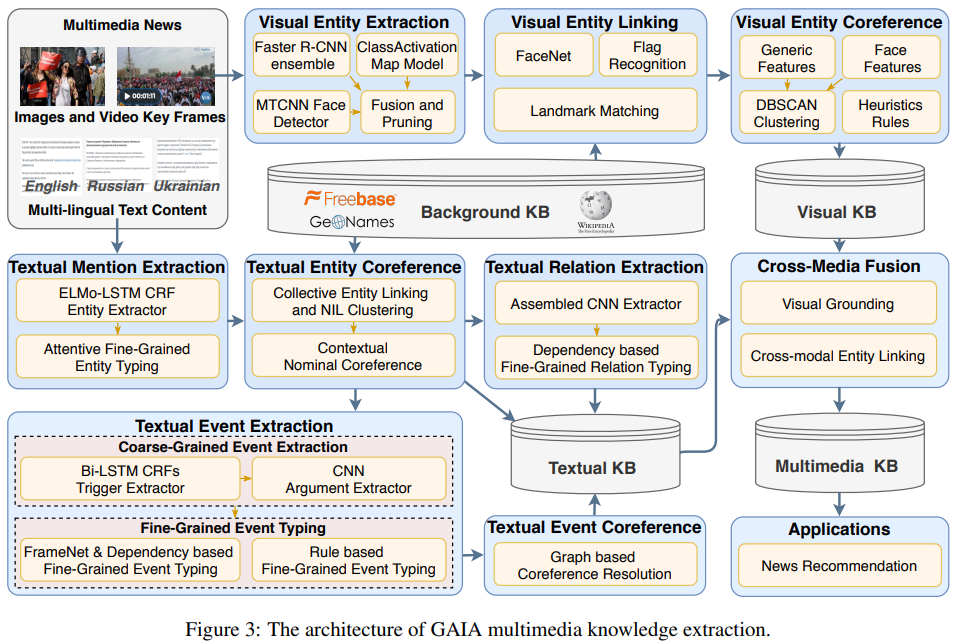
* 全面更新：指以更新后的全部数据为输入，从零开始构建知识图谱。这种方法比较简单，但资源消耗大，而且需要耗费大量人力资源进行系统维护；
* 增量更新：以当前新增数据为输入，向现有知识图谱中添加新增知识。这种方式资源消耗小，但目前仍需要大量人工干预（定义规则等），因此实施起来十分困难。



第二篇：『多模态』知识图谱构建

# GAIA: A Fine-grained Multimedia Knowledge Extraction System

<https://aclanthology.org/2020.acl-demos.11/>



The key point is that “information/knowledge” should also include “event”. Thus,

Information / Knowledge:

* Entity
* Relation
* Attribute
* Event

# Information from audio

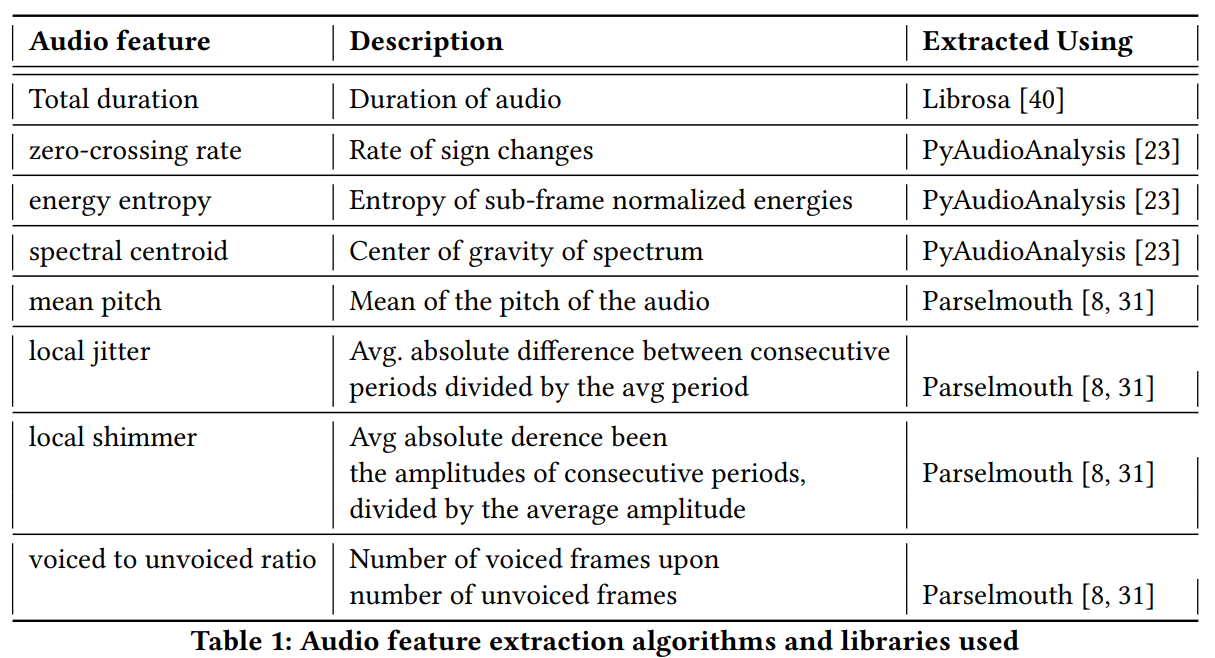
***What all do audio transformer models hear? Probing Acoustic Representations for language delivery and its structure***

<https://arxiv.org/pdf/2101.00387.pdf>



Thus, we learn that the information we obtain from audio data can be categorized into two parts:

* Audio-featured information
* Text-featured information





第三篇：架构图模板  
-基于多模态知识图谱的系统设计

