



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

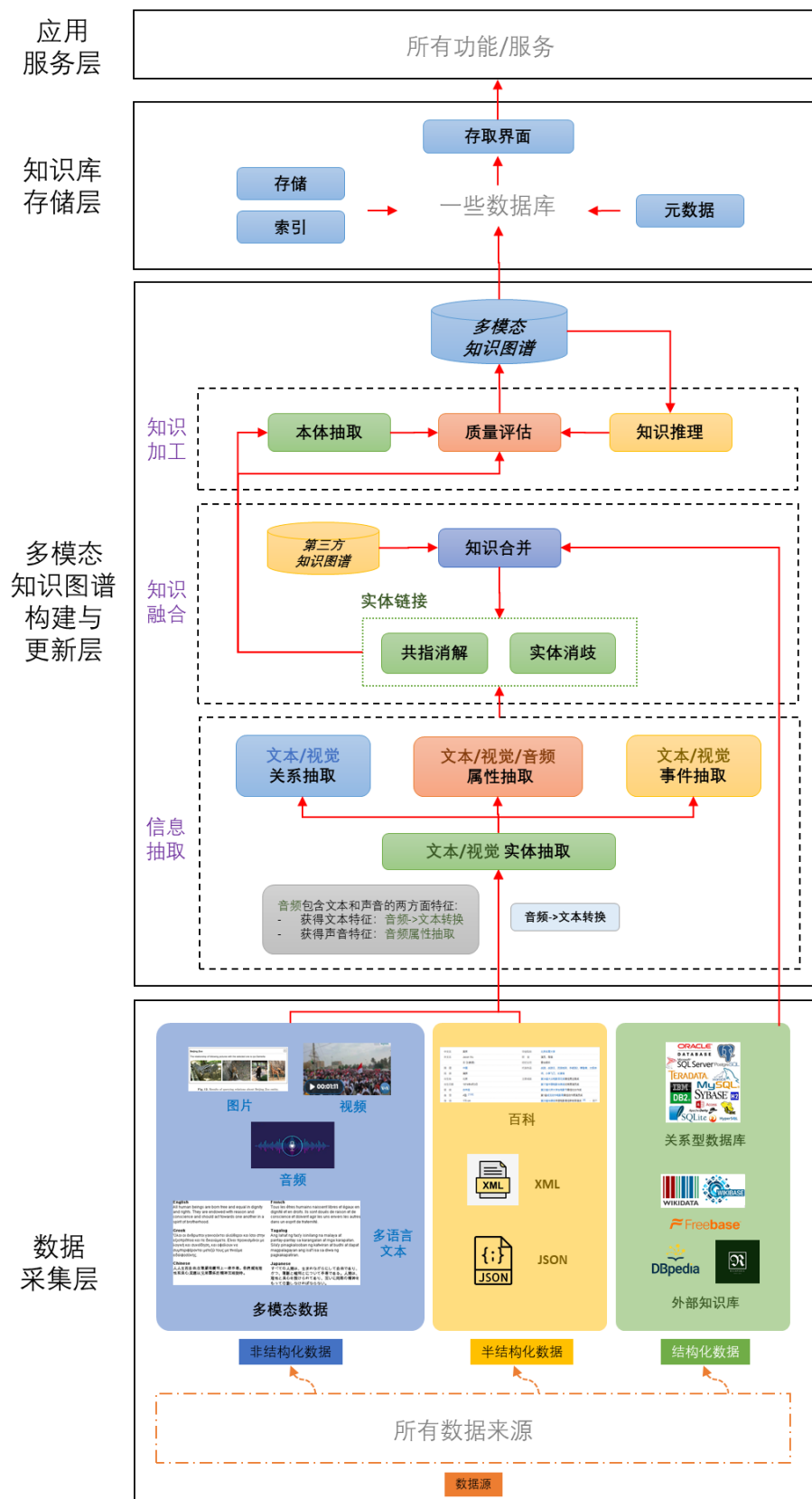
目录

架构图模板 -基于多模态知识图谱的系统设计	3
第一篇：知识图谱的构建与存储	4
1 知识图谱简介	4
1.1 广泛应用于各领域	4
1.2 构建技术分类	4
1.3 “实体-关系-实体”三元组	5
2 数据类型和存储方式	6
2.1 三类数据	6
2.2 存储方式：RDF（资源描述框架），图数据库	7
3 知识图谱的架构	8
3.1 逻辑架构	8
3.2 技术架构	8
4 构建技术	9
4.1 知识抽取	10
4.1.1 实体抽取	10
4.1.2 关系抽取	11
4.1.3 属性抽取	11
4.2 知识融合	12
4.2.1 实体链接	13
4.2.2 知识融合	14
4.3 知识加工	15
4.3.1 本体抽取	16
4.3.2 知识推理	17

4.3.3	质量评估.....	18
4.4	知识更新.....	18
第二篇：『多模态』知识图谱构建		20
1	GAIA: A Fine-grained Multimedia Knowledge Extraction System.....	20
2	Information from audio.....	21
第三篇：架构图模板 -基于多模态知识图谱的系统设计		23

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>



1 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。

1.1 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。

1.2 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。

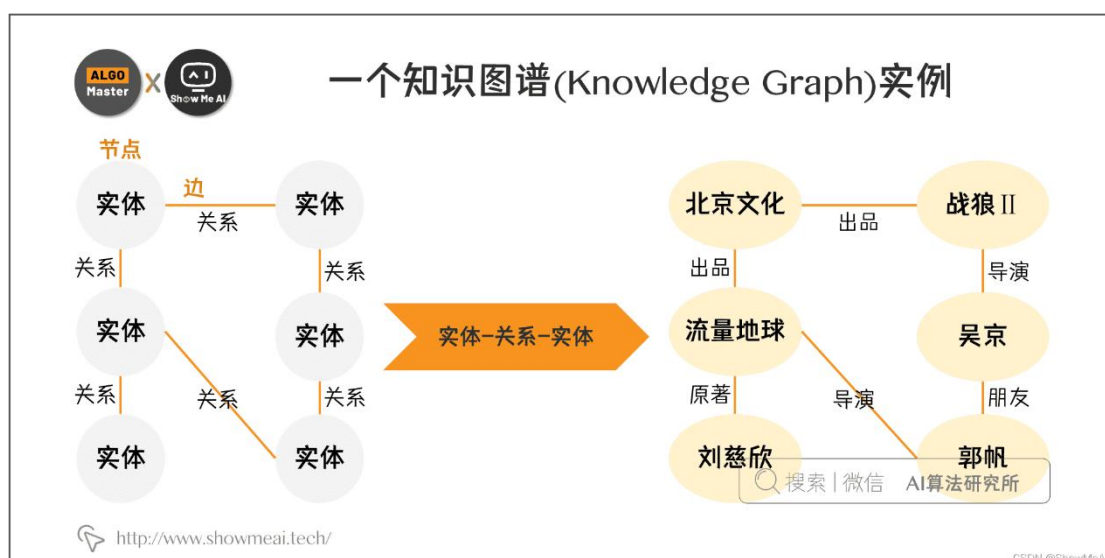


***本体**：在计算机科学和信息科学中，**本体**是指对**概念、数据和实体**之间的**类别、属性和关系**的表示、命名和定义，这些概念、数据和**实体**构成了一个、大量或所有的**论域^[1]**。

1.3 “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。




2 数据类型和存储方式

2.1 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



知识图谱 | 3种数据类型 & 2种存储方式

结构化数据

- 关系型数据库
- 链接数据

半结构化数据

- XML
- JSON
- 百科

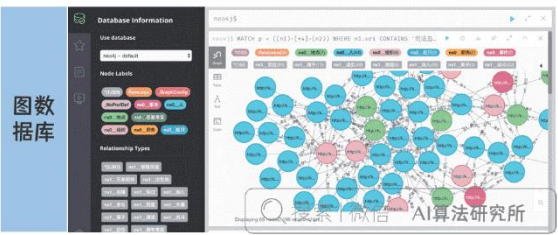
非结构化数据

- 图片
- 音频
- 视频

RDF

```
<RDF>
<Description about="http://www.showmeai.tech/RDF">
  <author>HanXinzi</author>
  <homepage> http://www.showmeai.tech </homepage>
</Description>
</RDF>
```

图数据库



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

CSDN @ShowMeAI

典型的半结构化数据样例如下：



知识图谱的数据类型示例 | 半结构化数据

中文名	吴京	毕业院校	北京体育大学
外文名	Jason Wu	职业	演员、导演
	오강(韩语)	经纪公司	音华娱乐
国籍	中国	代表作品	战狼、战狼II、流浪地球、杀破狼2、攀登者、太极宗师、小李飞刀、长津湖
民族	满族	主要成就	第34届大众电影百花奖最佳男主角奖
出生地	北京		第17届中国电影华表奖优秀男演员奖
出生日期	1974年4月3日		第22届北京大学生电影节最佳处女作奖
星座	白羊座		第1届成龙动作电影周最佳动作男演员奖
血型	A型 [130]		第20届华鼎奖年度电影最佳新锐导演奖 [4] 展开
身高	175 cm	配偶	谢楠
数据来源：百度百科		<input type="text" value="搜索"/> 微信 AI算法研究所	

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

CSDN @ShowMeAI

2.2 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>
```

2. 另一种方法是使用**图数据库**来进行存储，常用的有 Neo4j 等。

知识图谱的**存储方式**示例 | 图数据库

Neo4j

Neptune

TigerGraph

JanusGraph

ArangoDB

搜索 | 微信 AI算法研究所

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

CSDN @ShowMeAI

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？

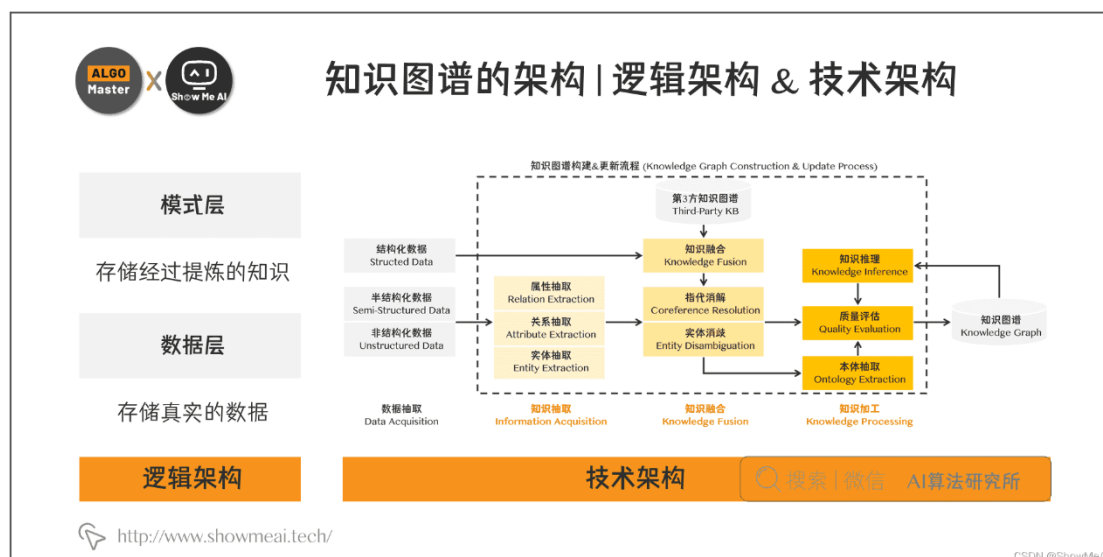
对，从技术上来说，用关系数据库来存储知识图谱（尤其是简单结构的知识图谱），是完全没问题的。但**一旦知识图谱变复杂**，用传统的「关系数据存储」，**查询效率**会显著低于「图数据库」。在一些涉及到 **2,3 度的关联查询**场景，**图数据库**能把查询效率**提升几千倍甚至几百万倍**。

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。

3 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



3.1 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

3.2 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。

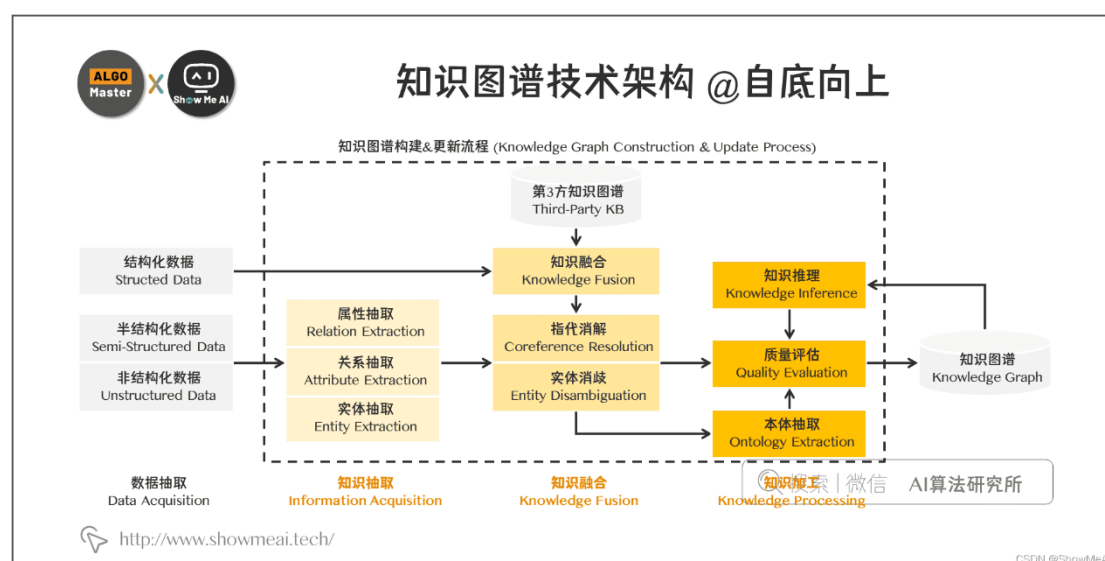
- 首先，我们有一大堆的数据，这些数据可能是结构化的、非结构化的以及半结构化的；
- 然后，我们基于这些数据来构建知识图谱，这一步主要是通过一系列自动化或半自动化的技术手段，来从原始数据中**提取出知识要素**，即**一堆实体、关系和属性**，并将其**存入我们的知识库的模式层和数据层**。

4 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。

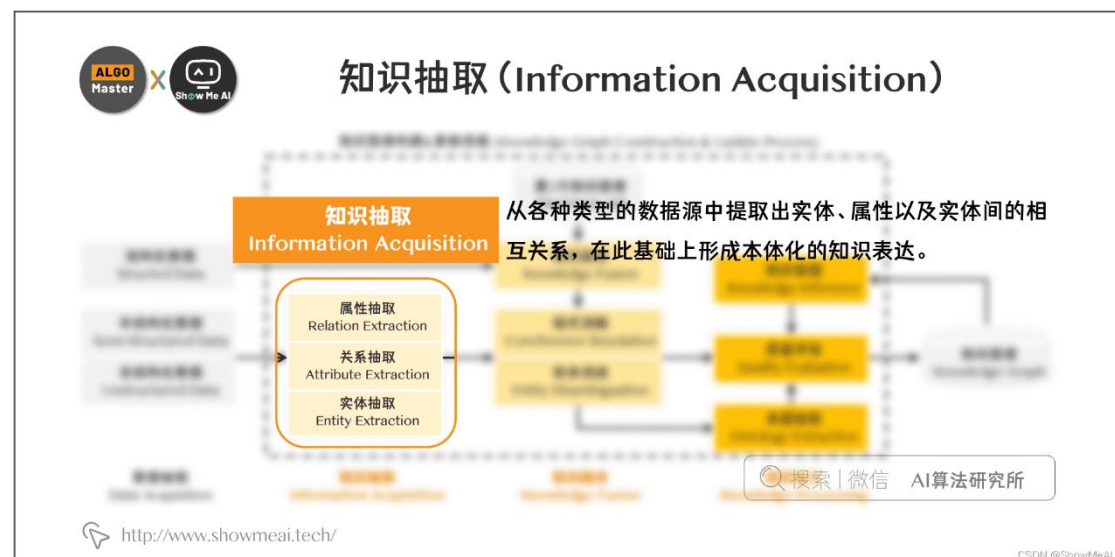


下面我们依次对每一个步骤进行介绍。

4.1 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**。



4.1.1 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。

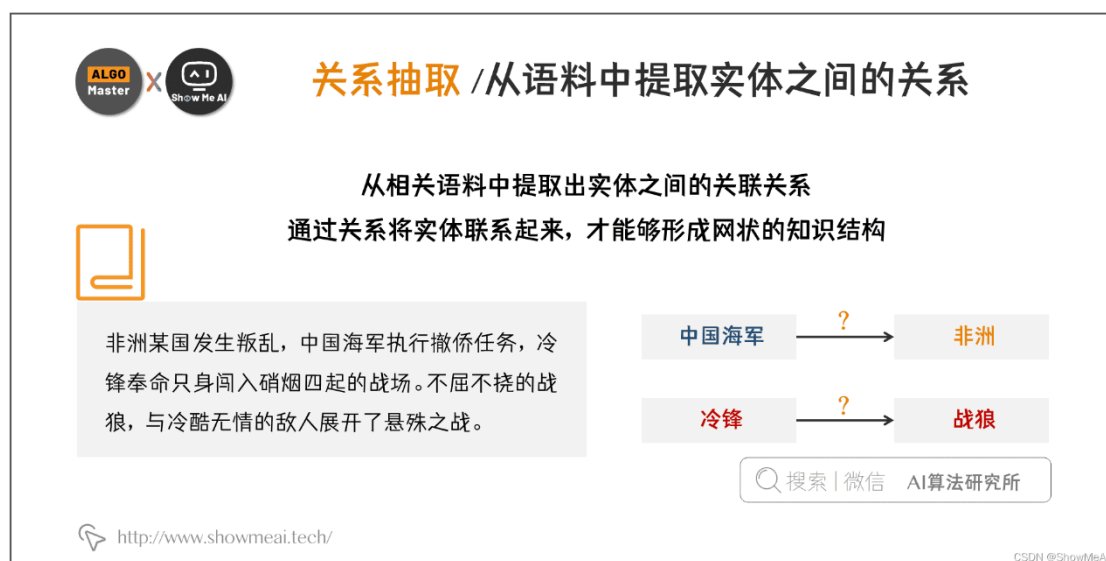
下图中，通过实体抽取我们可以从其中抽取四个实体：“非洲”、“中国海军”、“冷锋”、“战狼”。



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。

4.1.2 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。

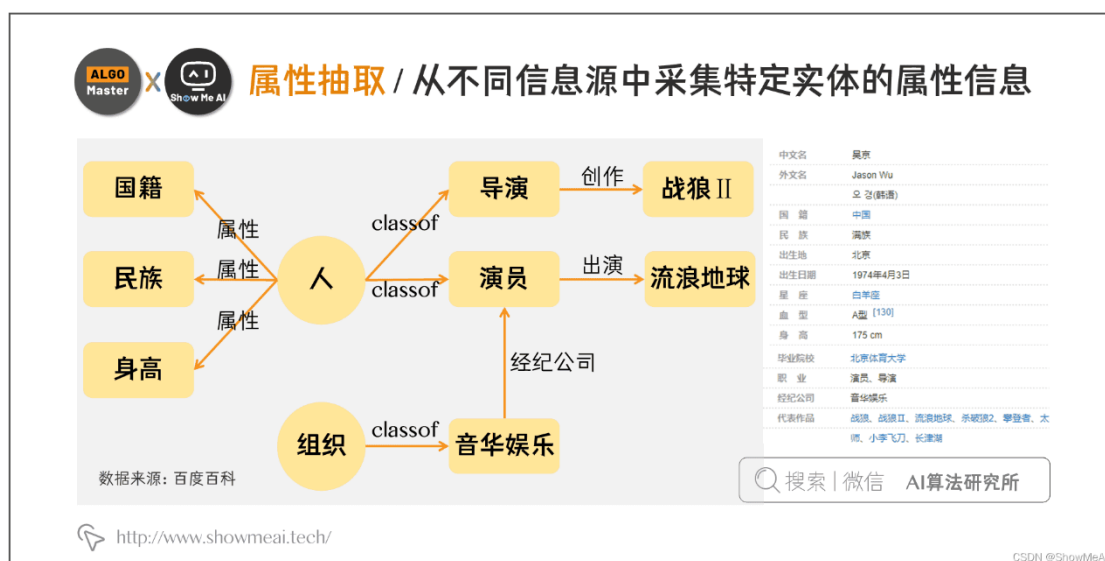
研究重点转向**半监督和无监督**。

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。

4.1.3 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：

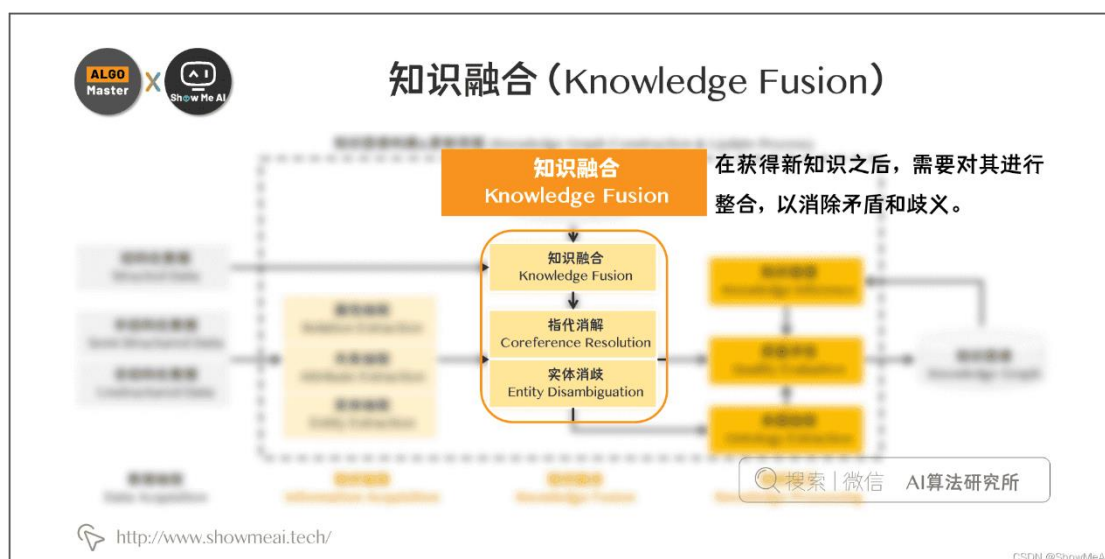
- 将实体的属性视作实体与属性值之间的一种**名词性关系**，将属性抽取任务转化为关系抽取任务。
- 基于**规则和启发式算法**，抽取**结构化数据**。
- 基于**百科类网站的半结构化数据**，通过自动抽取生成**训练语料**，用于训练实体属性**标注模型**，然后将其**应用于对非结构化数据的实体属性抽取**。
- 采用**数据挖掘**的方法**直接**从文本中挖掘实体属性和属性值之间的**关系模式**，据此实现对属性名和属性值在文本中的定位。

4.2 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。

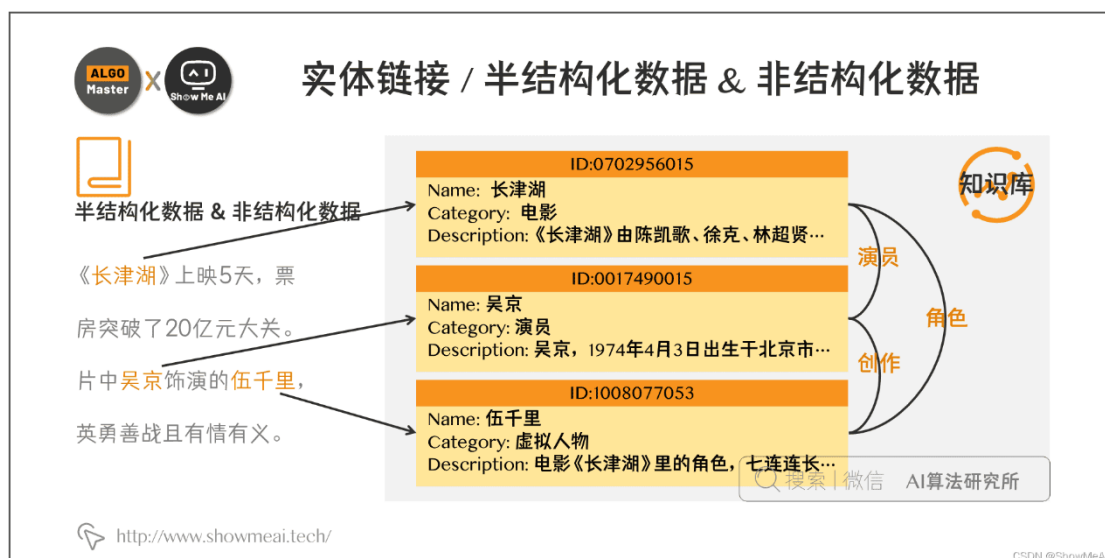
也就是说，**拼图碎片（信息）之间的关系是扁平化的，缺乏层次性和逻辑性**；**拼图（知识）中还存在大量冗余和错误的拼图碎片（信息）**。那么如何解决这一问题，就是在**知识融合**这一步里我们需要做的了。

知识融合包括 2 部分内容：**实体链接、知识合并**。



4.2.1 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**。

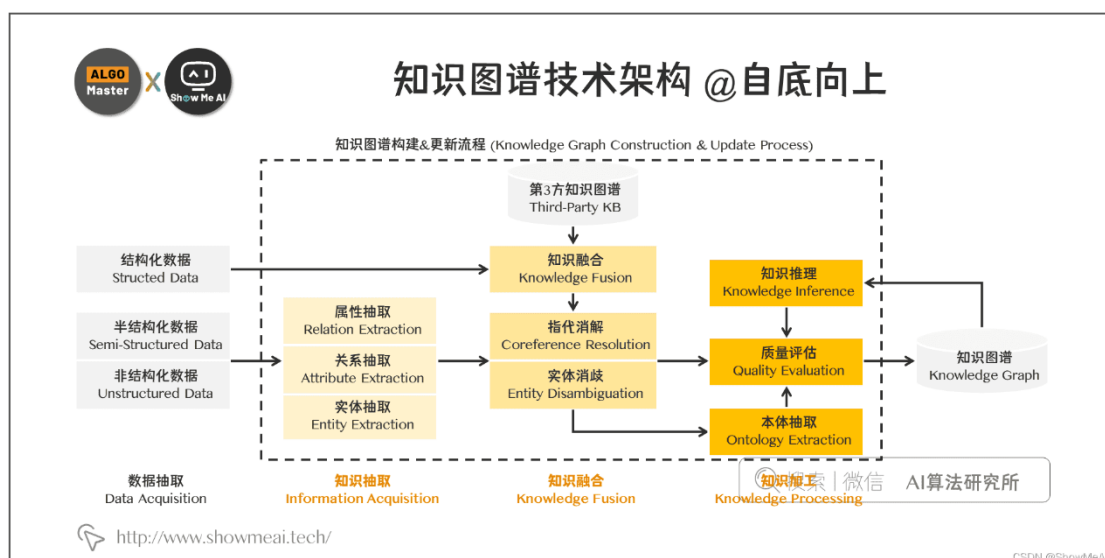
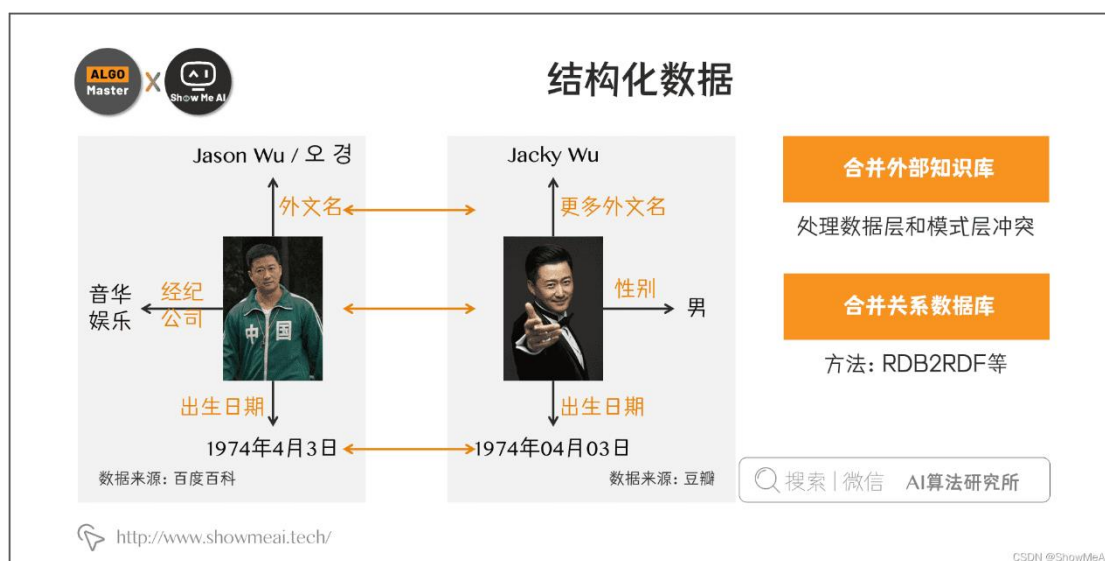
4.2.2 知识融合

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**：

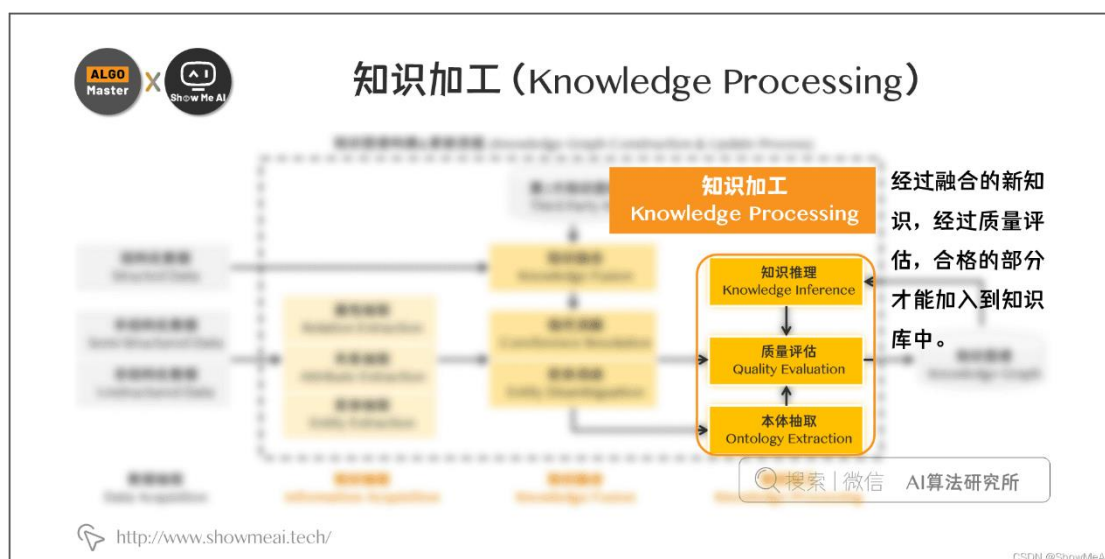
1. **合并外部知识库**，主要处理**数据层和模式层的冲突**；
2. **合并关系数据库**，有 **RDB2RDF** 等方法。



4.3 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**。



4.3.1 本体抽取

本体 (ontology) 是指工的概念集合、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』**

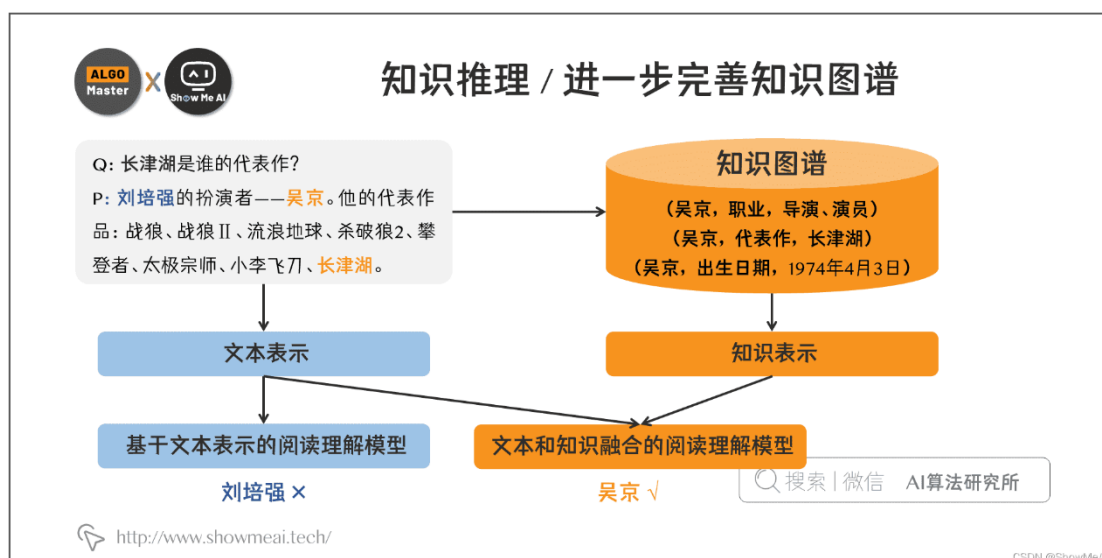
当三步结束后，这个知识图谱可能就会明白，“战狼 2 和流浪地球，是电影这个实体下的细分实体。它们和北京文化这家公司并不是一类”。

4.3.2 知识推理

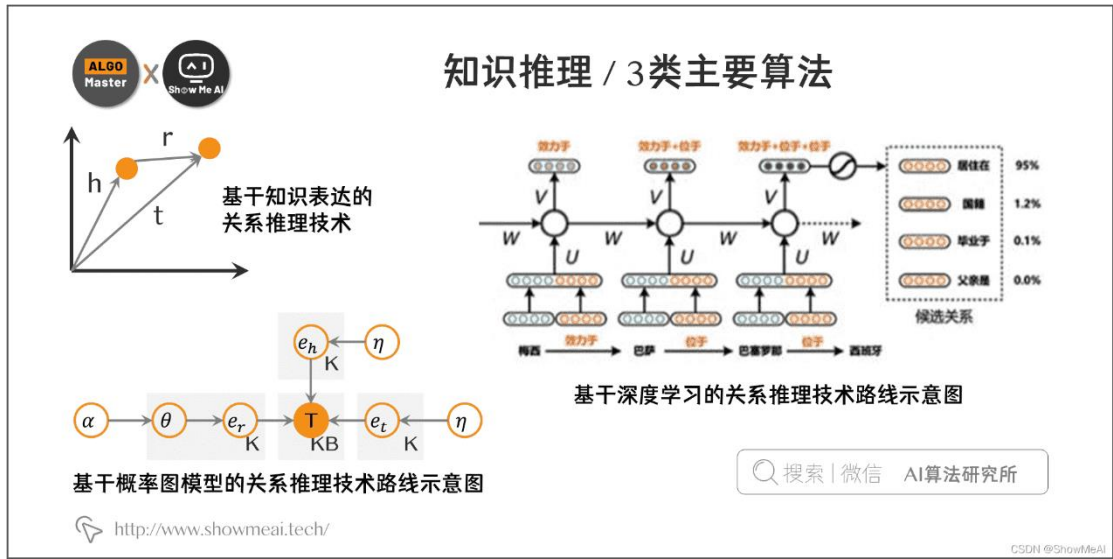
在我们完成了本体构建这一步之后，一个知识图谱的雏形便已经搭建好了。但可能在这个时候，知识图谱之间大多数关系都是残缺的，缺失值非常严重，那么这个时候，我们就可以使用知识推理技术，去完成进一步的知识发现。

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。



4.3.3 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。

4.4 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。

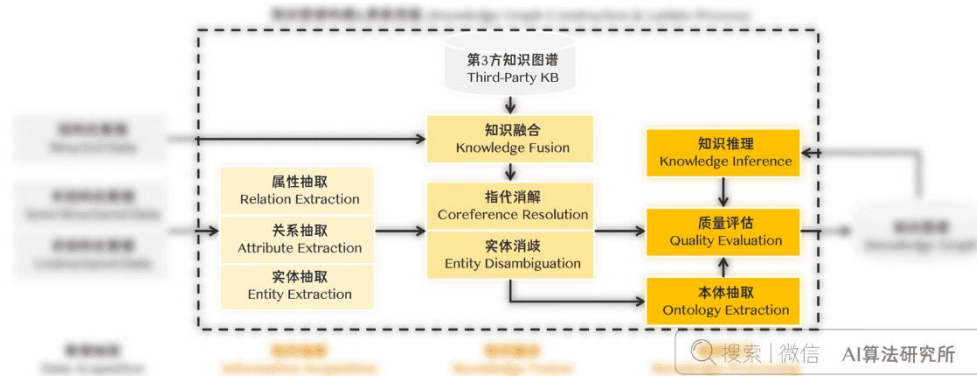
- **概念层的更新**：新增数据后获得了**新的概念**，需要自动将新的概念添加到**知识库的概念层**中。
- **数据层的更新**：主要是**新增或更新实体、关系、属性值**，对数据层进行更新需要考虑数据源的可靠性、数据的一致性（是否存在矛盾或冗杂等问题）等可靠数据源，并选择在各数据源中出现频率高的事实和属性加入知识库。

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。



知识图谱构建 & 更新流程



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

CSDN @ShowMeAI

第二篇：『多模态』知识图谱构建

1 GAIA: A Fine-grained Multimedia Knowledge Extraction System

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

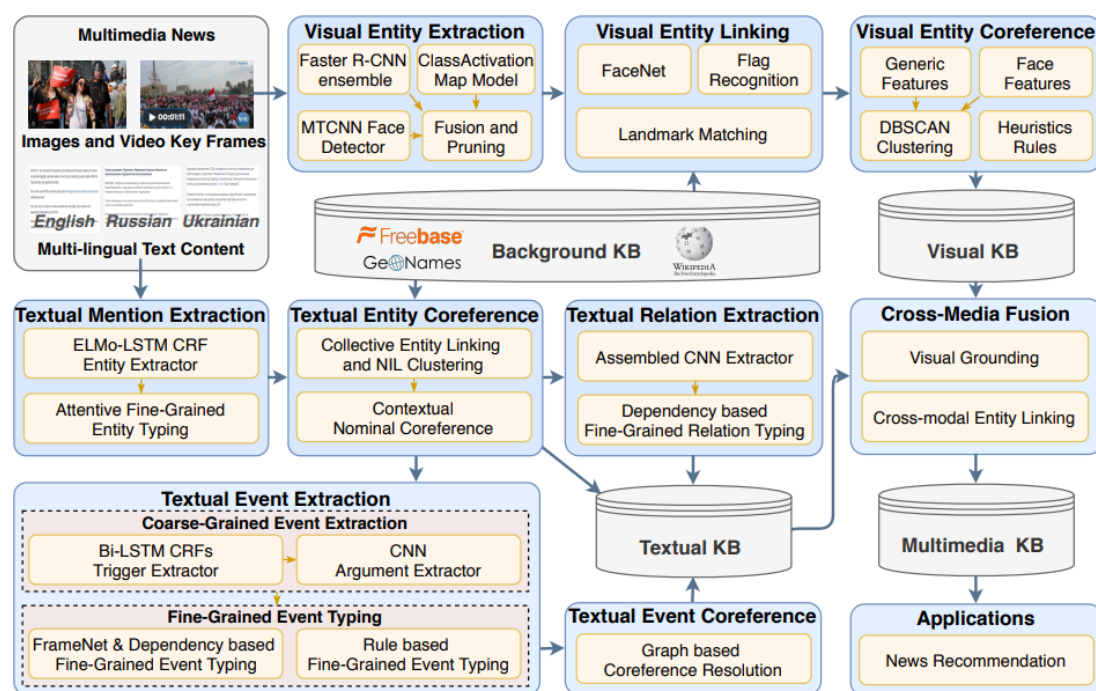


Figure 3: The architecture of GAIA multimedia knowledge extraction.

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

Information / Knowledge:

- Entity
- Relation
- Attribute
- Event

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

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

3.2 Feature Overview

We test the audio transformer models on the following speech features: audio features (§4.1), fluency features (§4.2), and pronunciation features (§4.3). Since spoken language can be considered as a combination of words (*what* was spoken), and language delivery (*how* it was spoken), we probe audio transformer models for both speech and text knowledge. For comparing on textual representational capacity, we extract text features from the original transcripts of all the audio datasets considered (§5). A detailed description of all features extracted and their methodology of extraction is given in Section 4 (audio features) and Section 5 (text features).

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

- Audio-featured information
- Text-featured information

Audio feature	Description	Extracted Using
Total duration	Duration of audio	Librosa [40]
zero-crossing rate	Rate of sign changes	PyAudioAnalysis [23]
energy entropy	Entropy of sub-frame normalized energies	PyAudioAnalysis [23]
spectral centroid	Center of gravity of spectrum	PyAudioAnalysis [23]
mean pitch	Mean of the pitch of the audio	Parselmouth [8, 31]
local jitter	Avg. absolute difference between consecutive periods divided by the avg period	Parselmouth [8, 31]
local shimmer	Avg absolute derence been the amplitudes of consecutive periods, divided by the average amplitude	Parselmouth [8, 31]
voiced to unvoiced ratio	Number of voiced frames upon number of unvoiced frames	Parselmouth [8, 31]

Table 1: Audio feature extraction algorithms and libraries used

Text feature	Description
Surface Features	
Unique word count	Total count of unique words(Ignore words of length 3 or smaller)
Word Complexity	Sum of word complexities for all words in text given by annotators
Semantic Features	
Total adjectives	Total count of adjectives
Total adverbs	Total count of adverbs
Total nouns	Total count of nouns
Total verbs	Total count of verbs
Total pronouns	Total count of pronouns
Total conjunction	Total count of conjunction
Total conjunction	Total count of conjunction
Number of subject	Total count of subject
Number of Object	Total count of direct objects
Tense	Classification of main clause verb into present or past tense
Syntax Feature	
Depth of syntax tree	Depth of syntax tree of the text

Table 4: Text feature extraction algorithms extracted using nltk and numpy libraries

第三篇：架构图模板

-基于多模态知识图谱的系统设计

