知识图谱构建流程

1. 知识获取（信息抽取）

知识获取的主要目的就是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。而其中的关键问题是：如何从异构数据源中自动抽取信息得到候选指示单元？

1. 结构化数据

结构化数据主要是针对关系型数据库的数据，数据结构清晰。

1. 半结构化数据

半结构化数据是指一些具有一定数据结构的，但需要进一步提取整理的数据。比如百科的数据、网页中的数据。

针对半结构化数据，可以采用手工方法、包装器归纳、自动抽取等方式进行处理。

|  |  |  |  |
| --- | --- | --- | --- |
|  | 手工方法 | 包装器归纳 | 自动抽取 |
| 优点 | 1. 对于任何一个网页都是通用的，简单快捷； 2. 能抽取到用户感兴趣的数据 | 1. 需要人工标注数据集 2. 能够抽取到用户感兴趣的数据 3. 可以运用到大规模网站的信息抽取 | 1. 无监督的方法，无需人工进行数据的标准 2. 可以运用到大规模网站的信息抽取 |
| 缺点 | 1. 需要对网页数据进行标注，耗费大量的人力 2. 维护成本高 3. 无法处理大量站点的情况 | 1. 可维护性比较差 2. 需要投入大量的人力去做标注 | 1. 需要相似的网页作为输入 2. 抽取的内容可能打不到预测的情况。会抽取一些无关信息 |

1. 非结构化数据
   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。

实体抽取的研究历史主要是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。

* 1. 属性抽取

研究关系抽取技术，就是研究如何解决从文本语料中抽取实体间的关系。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（节点），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（边），才能将多个实体或概念联系起来，形成网状的知识结构。

* 1. 关系抽取

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。如果把实体的属性值看作是一种特殊的实体，那么属性抽取实际上也是一种关系抽取。

1. 知识融合
2. 实体链接
3. 知识合并

由于知识图谱的数据源可能有多个，所以在获取到的数据中可能存在着同名实体代表着不同含义或者知识库中存在多个命名实体表示同一含义的的情况。

1. 知识加工
2. 本体构建
   1. 实体并列关系相似度计算
   2. 实体上下位关系抽取
   3. 本体的生成

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。

1. 知识推理

知识推理，就是根据已有的数据模型和数据，依据推理规则，获取新的知识或者结论，新的知识或结论应该是满足语义的。

1. 质量评估

质量评估，就是对最后的结果数据进行评估，将合格的数据放入知识图谱中。质量评估的方法，根据所构建的知识图谱的不同，对数据要求的差异而有所差别。总的目的是要获得合乎要求的知识图谱数据，要求的标准根据具体情况确定。

1. 知识更新
2. 全量更新

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。

1. 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。