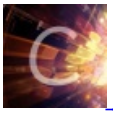
华为云专家讲述知识图谱构建流程及方法

原创

华为云开发者社区 于 2020-10-12 11:19:46 发布 2063 收藏 7
分类专栏： 技术交流 文章标签： 知识图谱 华为云 数据分析 信息抽取 训练模型



摘要：随着AI技术的发展和普及，当今社会已经进入了智能化时代。与以往不同的是，在这一波浪潮中，企业不仅是向数字化转型，更是向知识化转型。那么，如何助力企业破解智能化知识挖掘和管理难题，实现知识化转型？

华为云自然语言处理技术专家郑毅在《企业级知识计算平台的技术解读和案例实践》分享中，讲述了华为云知识计算平台及相关技术、知识图谱构建流程及方法，以及知识计算行业案例。本文主要讲述“知识图谱构建流程及方法”，让我们先睹为快。

一、 什么是知识图谱？

知识图谱是由实体、关系和属性组成的一种数据结构。以下图为例，“刘德华“是一个人物类型的实体，“刘德华”有自己的身高、国籍等信息，这些信息便称之为实体的属性。

同样，“无间道”是一个电影类型的实体。我们知道“刘德华”是“无间道”这部电影的主演，所以“刘德华”与“无间道”之间有“主演”关系。通过实体、关系、属性，就能够把我们人可以理解的知识有效地组织起来。知识图谱的构建与应用涉及数据库、自然语言处理（NLP）和语义网络等技术。

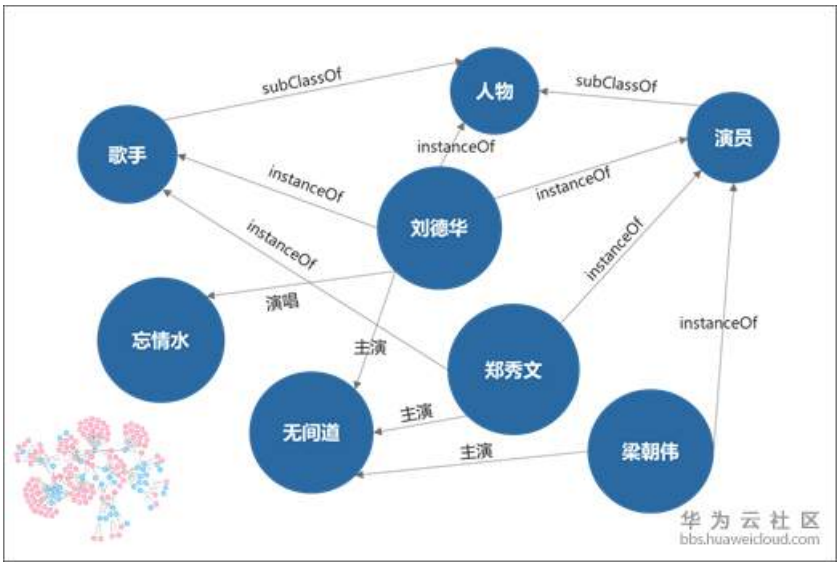


图1 知识图谱示例

通用知识图谱or行业知识图谱？

按照知识图谱的用途，知识图谱可分为通用知识图谱和行业知识图谱。通用知识图谱侧重构建常识性的知识，并用于搜索引擎和推荐系统等。行业知识图谱（也可称企业知识图谱）主要面向企业业务，通过构建不同行业、企业的知识图谱，对企业内部提供知识化服务。华为云知识图谱服务可用于以上两类知识图谱的构建、管理和服务，更侧重面向企业知识图谱。

二、 如何构建知识图谱？

知识图谱构建主要分为自顶向下(top-down)与自底向上(bottom-up)两种构建方式。自顶向下构建方式需要先定义好本体（Ontology或称为Schema），再基于输入数据完成信息抽取到图谱构建的过程。该方法更适用于专业知识方面图谱的构建，比如企业知识图谱，面向领域专业用户使用。自底向上构建方式则是从开放的Open Linked Data中抽取置信度高的知识，或从非结构化文本中抽取知识，完成知识图谱的构建。该方式更适用于常识性的知识，比如人名、机构名等通用知识图谱的构建。本文侧重介绍自顶向下构建方式的相关流程和技术，并用于构建企业知识图谱。

目前业界暂无知识图谱云服务，也没有统一标准的自顶向下构建流程。当前业界主流的知识图谱构建方式是基于企业内部数据、公开数据，图谱服务商以解决方案形式帮助客户定制构建知识图谱。这样的方式无疑成本非常高并且效率很低，通常需要很长的周期才能完成。同时，企业没有参与感，图谱构建也可能存在很大偏差，难以用于实际业务中。

站在用户角度，我们通过抽象知识图谱构建流程及相关技术，推出华为云知识图谱云服务（图2），为不同行业、不同企业提供快速构建知识图谱能力的平台，赋能大中小型企业构建属于自己的知识图谱。

图2 华为云知识图谱云服务

华为云知识图谱云服务提供流水线式图谱构建能力，将图谱构建抽象为如下基本流程：本体构建、数据源配置、信息抽取、知识映射以及知识融合。

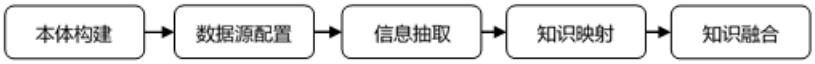


图3 知识图谱构建基本流程

进一步通过将每一个流程模块抽象成插件形式，并通过组合配置生成图谱构建任务。面向不同的行业和领域，只需要修改插件配置即可完成企业知识图谱的构建。同时，基于流水线设计，知识图谱云服务可以在只修改数据源的前提下完成知识图谱的更新操作，非常适用于需要频繁更新的知识图谱。

2.1 如何构建知识图谱的本体？

知识图谱构建的第一步需要完成图谱本体（Ontology）的设计和构建。本体是图谱的模型，是对构成图谱的数据的一种模式约束。对于企业知识图谱的构建，一般是由垂直领域的行业专家和知识图谱专家合作完成。

本体的构建和设计对于知识图谱的构建至关重要。可以通过梳理领域知识、术语词典、专家的人工经验等作为本体构建的基础，结合知识图谱的应用场景来完善图谱的构建，最终获得实体类别、类别之间的关系、实体包含的属性定义。华为云知识图谱云服务提供图形化本体设计工具，可以通过拖拽编辑灵活完成企业知识图谱本体的构建。

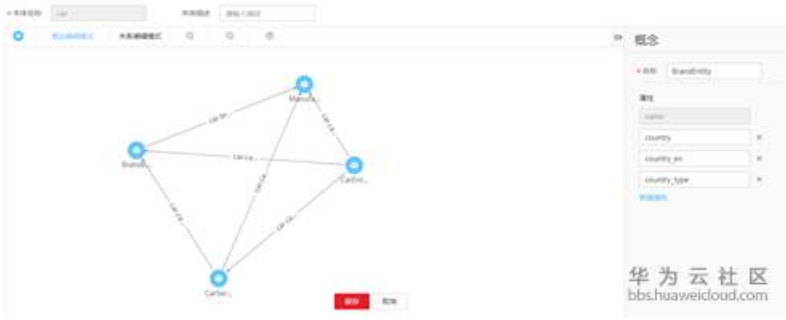


图4 华为云知识图谱云服务-本体设计界面

2.2 如何配置数据源？需要做哪些准备

在配置数据源之前，需要将不同类型、不同格式的数据进行初步的整理。比如：针对本地非电子化文档，需要先进行扫描电子化，结合OCR等技术将扫描件转换成文本文档。再比如：针对本地电子化文档，需要将本地文档按文档类型、格式进行归档解析整理成规范的格式，或者针对网络资源，需要根据网站特点，开发相应的爬虫，对数据进行爬取，并存储到本地数据库等等。还有一些第三方资源，需要获取相应的数据访问接口，并通过接口获取相应数据。

整理好的数据上传到华为云OBS对象存储服务后，知识图谱云服务就可以进行数据源的配置，包括指定格式的针对结构化数据和非结构化文本的配置等。

2.3 什么是信息抽取？怎样抽取？

信息抽取的目的是根据不同的数据源、不同的数据格式，完成实体、属性、关系这种知识的抽取。这是知识图谱构建流程中非常关键的一环，信息抽取的质量决定了知识图谱的质量。实体之间的关系以及实体的属性值，都可以用三元组（主语、谓词、宾语）来表示，所以信息抽取又可以简单叫做三元组抽取。

华为云知识图谱云服务支持结构化Key-Value格式和非结构化文本的三元组抽取。针对结构化数据，可以通过配置预置函数的组合，完成字段的处理。与之对应的，针对非结构化文本，云服务提供算法模型抽取能力，支持业界前沿的基于机器阅读理解（Machine Reading Comprehension，MRC）的三元组抽取方法，通过使用多轮对话的思想进行三元组抽取，先抽取主语（Subject），然后根据抽取结果和候选谓词对应的模板构造问句抽取宾语（Object），最终组成（主语，谓词，宾语）三元组。该框架模型效果可以达到当前业界最好水平（state-of-the-art）。华为云知识图谱服务支持基于该算法的模型训练、预测以及管理功能，同时以插件形式完成流水线中信息抽取部分。

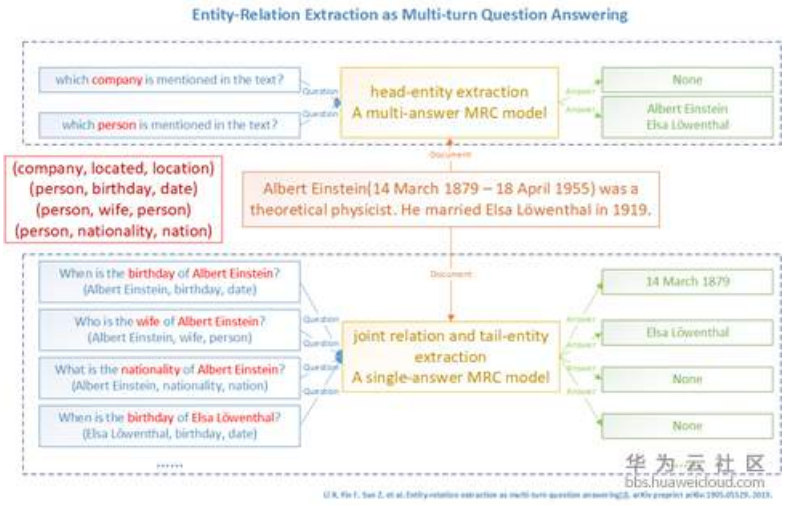


图5 基于机器阅读理解（MRC）的三元组抽取方法

信息抽取中模型训练推理功能是基于华为云-ModelArts AI计算平台完成的，该平台提供高效的AI计算、模型训练、推理及部署能力，同时为了方便训练三元组抽取模型，额外提供三元组标注工具，用户可以基于该工具快速获得训练数据，完成信息抽取以及知识图谱构建工作。

图6 三元组标注工具示例

2.4 知识融合是如何完成的？

所谓知识融合，就是对多个数据源进行知识抽取后的大量三元组数据进行对齐合并。举个例子：百度百科有明星刘德华，互动百科有明星刘德华，我们构建的知识图谱不能有两个明星刘德华吧？这时候就需要把他们识别出来放在一起，然后合并成一个实体，这就是实体的对齐以及知识的融合。

这其中关键的问题是怎样高效的完成实体对齐，技术路线基本可以分为两类：基于实体属性相似度的框架、基于联合表征的深度学习框架。考虑到基于联合表征的深度学习框架依赖大量标注数据，并且模型与行业及数据强相关，无法提供很好的通用化能力，因此，华为云知识图谱服务当前支持基于实体属性相似度的框架，可以通过定义相似度度量及组合，完成实体对齐以及知识融合。

除此之外，华为云知识图谱云服务还提供图谱可视化服务，可以直观地观察分析实体及关系。

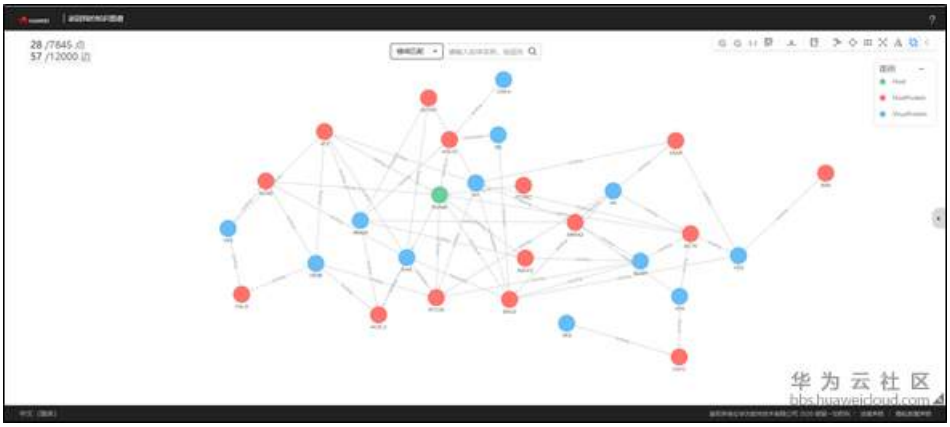


图7 病毒蛋白知识图谱可视化示例

三、 知识图谱需要怎样的存储方式？

经过知识图谱构建，我们现在已经有了大量的三元组知识。那么要怎样来存储这些三元组知识呢？

最直接的方式是使用表格式的存储方式，如关系型数据表，三元组以三列数据或多列数据的形式存储。这种方法在图谱规模比较小的时候是可行的，但是如果图谱规模变大了，是否依然可行呢？举个例子，假使我们有了娱乐明星+电影这样一个娱乐图谱，其中包括了大量的明星人物、电影以及他们之间的关系。如果想查询“刘德华和梁朝伟共同演过的电影中，年龄最大的导演是谁？“，就需要对关系型数据库中知识图谱结果表做2-3次自连接操作，如果三元组的数量是千万、亿、十亿规模的话，显而易见，这样的查询效率极低，基本不可行。

华为云知识图谱服务采用的是业界主流的图数据库方式存储知识图谱，直接把数据或知识图谱以图的形式存储，可以非常高效地完成多跳关系、属性的查询。具体的，我们使用华为云图引擎服务，包括图存储、图计算一体的架构设计，不仅可以提供高效的查询性能，同时也可以提供多种预置的图深度学习算法，使用起来非常方便，欢迎大家前来试用。

图8 华为云图引擎服务产品优势

四、 华为云知识计算案例介绍

中国石油基于华为云知识计算服务的知识建模、油气图谱构建、图谱存储、自然语言处理、机器学习等能力构建了业界首个油气知识计算平台。以油气勘探开发数据为基础，通过知识计算技术的应用，为油气勘探开发增储上产、降本增效提供智能辅助和决策。

图9 油气知识计算的价值和意义

华为知识计算解决方案提供丰富的知识应用，从解决企业痛点、提升企业效率、提供知识化服务的角度全面赋能企业，体现了知识计算在各行业中的智能化价值，让各行业的企业可以快速、低成本、高效率地管理，通过应用企业知识、实现知识化转型，释放知识化带来的红利，全面提升企业在智能化时代的竞争力。

点击关注，第一时间了解华为云新鲜技术~