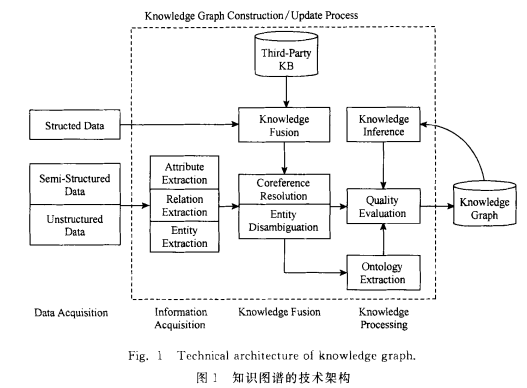
人物知识图谱构建

总体思路: 自顶向下的构建模式，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，加入到知识库中



1. 语料获取

数据来源 （1）百度百科数据 （2）维基百科数据

主要信息格式 （1）结构化信息 （2）非结构化信息

1. 信息抽取

主要从非结构化信息中抽取三元组

三元组内容如下：



1. 人物实体抽取

命名实体识别，从文本数据集中自动识别出命名实体，实体抽取的质量（准确率和召回率）对后续的知识获取效率和质量影响极大，因此是信息抽取中最为基础和关键的部分。人物实体感觉大部分被infobox定义完毕了。

待用模型：BiLSTM-CRF，BERT系列

1. 关系抽取（只抽人物之间的关系）

从相关语料中提取出实体之间的关联关系。人物知识图谱的关系有限，定义不到很多。实体关系抽取的模型看后续可不可以引进图谱构建中。

待用模型：MTB

1. 属性抽取

该过程可以分为两种:

1. 抽出的属性为非人物实体，建立实体和关系
2. 抽出的属性不是非人物实体，添加属性值

从相关语料中采集特定实体的属性信息。

由于有网页有infobox，因此需要先对其进行抽取

待用模型：MRC系列

补充：在抽取的过程中尽量利用结构化信息

1. 知识融合

此步需准备（1）基于结构化信息得到一个不完备的知识图谱，（2）以及加上信息抽取获得的信息。信息抽取中获得的信息可能会包含大量的冗余和错误，数据之间的关系也是扁平化的，缺乏层次性和逻辑性，此步主要做：

1. 归一化和标准化

主要利用规则进行属性值上单位等信息的统一，以及消除一部分概念的歧义

1. 实体链接

对于从文本中抽取得到的实体对象，将其链接到知识库中的对应的正确实体对象的操作，主要包括实体消岐和共指消解，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

待用模型：周沛的工作

1. 知识合并

将抽取出的信息和原有的知识库进行融合，此步可能涉及百度百科和维基百科数据的融合

1. 知识加工（可选）

知识融合可以消除实体指称项与实体对象之间的歧义，得到一系列的基本的事实表达。但是事实并不等于知识，还需要知识加工，此步主要做：

1. 本体构建

本体是同一领域内的不同主体之间进行交流的语义基础，即schema。

1. 知识推理

从知识库中已有的实体关系数据出发，经过计算机推理，建立实体间的新关联。例如已知（乾隆，父亲，雍正）和（雍正，父亲，康熙），可以得到（乾隆，祖父，康熙）或（康熙，孙子，乾隆）

1. 质量评估
2. 知识更新

知识库的迭代更新