中国近代历史人物知识图谱

王天笑 · 计算机科学与技术学院 软件工程 · 22121131

> 项目地址：http://www.zjuwtx.work/project/kg

1 引言

中国近代历史涌现了大量的杰出人物和事迹，史料文献丰富，构成了一个庞大的知识体系。本项目旨在收集挖掘中国近代历史人物信息，构建人物及其相关实体的知识图谱，为历史资料的检索和研究工作提供帮助。

2 知识图谱构建

2.1 数据来源

中国近现代历史人物信息主要来源于百度百科和历史记两个网站。通过 python scrapy 爬虫获取了近1300位人物的结构化数据，半结构化数据和文本数据。其中，结构化数据主要包含人物的姓名、字号、出生地、生卒年月等信息；半结构化数据包括人物间的关系、历史成就等；文本数据主要是人物的生平介绍、评论等，有网站负责编辑维护，语言描述和记录的史料不一定完全准确。

2.2 数据处理

2.2.1 半结构化数据处理

- 数据变换：从半结构化数据中提取信息，转化汇总成相同的格式，例如

> 著有《xxx》 -> 文学作品：《xxx》

- 数据清理：将明显不正确的信息删除，例如

> 出生年月：1766年。明显超出了近代范畴，可能是网站的链接信息出错

- 数据集成：整合两个数据源的数据，如果有不一致，直接删除

2.2.2 文本数据处理

项目尝试了基于语义角色标注和基于深度学习的实体关系抽取方法。

2.2.2.1 基于 LTP 语义角色标注的实体关系抽取

LTP（Language Technology Platform）是由哈尔滨工业大学开源的中文自然语言处理工具，用户可以使用这些工具对于中文文本进行分词、词性标注、句法分析等等工作。Github：https://github.com/HIT-SCIR/ltp

使用LTP提供的语义角色标注分解句子：

Python 复制代码

```python
from ltp import LTP
ltp = LTP()
seg, hidden = ltp.seg(["王俊昌于1943年2月加入中国共产党。"])
# seg: [['王俊昌', '于', '1943年', '2月', '加入', '中国', '共产党']]
srl = ltp.srl(hidden, keep_empty=False)
# srl: [[(4, [('A0', 0, 0), ('ARGM-TMP', 1, 3), ('A1', 5, 6)])]]
```

上述例句被分解为了中心语（动词：加入），主语A0（王俊昌），宾语A1（中国共产党）和时间状语ARGM-TMP（1943年2月）。通过构建基于语义角色标注的规则，可以从文本数据中提取符合规则的关系，准确度较高但规则构建依赖人工。

2.2.2.1 基于OpenUE的实体关系抽取

OpenUE 是一个轻量级知识图谱抽取工具，用于基于预训练语言模型的知识图谱抽取任务。Github：https://github.com/zjunlp/OpenUE

使用OpenUE工具包和默认ske数据集训练并执行抽取。在简单句子中准确率较高，但是在所有文本数据中的表现并不理想。原因可能是文本语句通常比较复杂，且句子间存在上下文关联的情况。例如主语缺失等。

2.2.2.3 基于OpenNRE的人物关系抽取

OpenNRE 是一个开源且可扩展的工具包，它提供了一个统一的框架来实现关系提取模型。项目尝试使用基于OpenNRE的中文人物关系抽取，Github：https://github.com/taorui-plus/OpenNRE

按照上述Github项目的描述训练模型并执行关系提取任务，结果同样在简单句型中表现良好，但在多数复杂句型中出现了遗漏和错误。

综上所述，出于准确度、史实正确性优先的考虑，项目最终使用了基于语义角色标注的实体关系抽取方法。

3 知识图谱存储

项目基于neo4j图数据库存储实体关系数据。

实体对象共3类：人物，组织（学校），成就（作品）。其中人物包含属性：名称、附加名称、出生地、出生日期、死亡日期、工作职责、名族、国籍（在华外籍人物）。

实体关系共3个大类：相关人物、毕业于、创作。相关人物可细分为7个子类，21个具体关系，如下所示

JSON 复制代码

```
  '相关人物-父子': ['相关人物-儿子', '相关人物-女儿', '相关人物-父亲', '相关人物-母亲'],
  '相关人物-祖孙': ['相关人物-孙子', '相关人物-孙女', '相关人物-爷爷', '相关人物-奶奶'],
  '相关人物-兄弟': ['相关人物-哥哥', '相关人物-妹妹', '相关人物-弟弟', '相关人物-姐姐'],
  '相关人物-婚配': ['相关人物-丈夫', '相关人物-妻子'],
  '相关人物-婿媳': ['相关人物-女婿', '相关人物-儿媳'],
  '相关人物-师生': ['相关人物-学生', '相关人物-老师'],
  '相关人物-其他': ['相关人物-战友', '相关人物-同学', '相关人物-好友'],
```

4 知识图谱应用

项目最终成果使用BS形式部署上云。后端打包为Docker镜像部署到阿里云ECI，前端部署到阿里云CDN。可以访问 http://www.zjuwtx.work/project/kg 查看。

4.1 人物检索

基本的人物检索功能，查看人物属性以及于其他实体间的关系。

钱钟书

id：886

label：人物

出生地：江苏无锡县

出生日期：1910年10月20日

名称：钱钟书

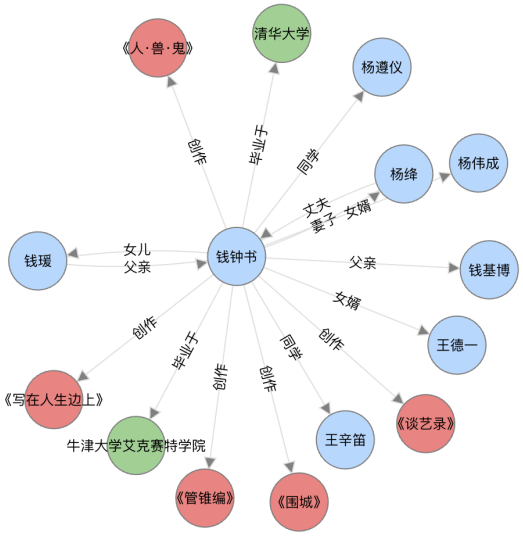
国籍：中国

工作职责：作家、文学研究家

死亡日期：1998年12月19日

民族：汉族

附加名称：钱仰先，字默存，号槐聚



4.2 图谱推理

基于规则的图谱推理，通过自定义Cypher脚本实现。包括关系推理和属性补全。

4.3 知识众包

考虑到数据来源有限，同时数据内容以及数据处理过程不可避免地会存在一些问题，导致了图谱知识的缺失和错误。项目提供了知识众包功能，所有用户可以快速提交新增、修改数据的请求，在审核通过后会合并到现有的知识图谱中。