每秒24格的真理

——电影人物知识图谱

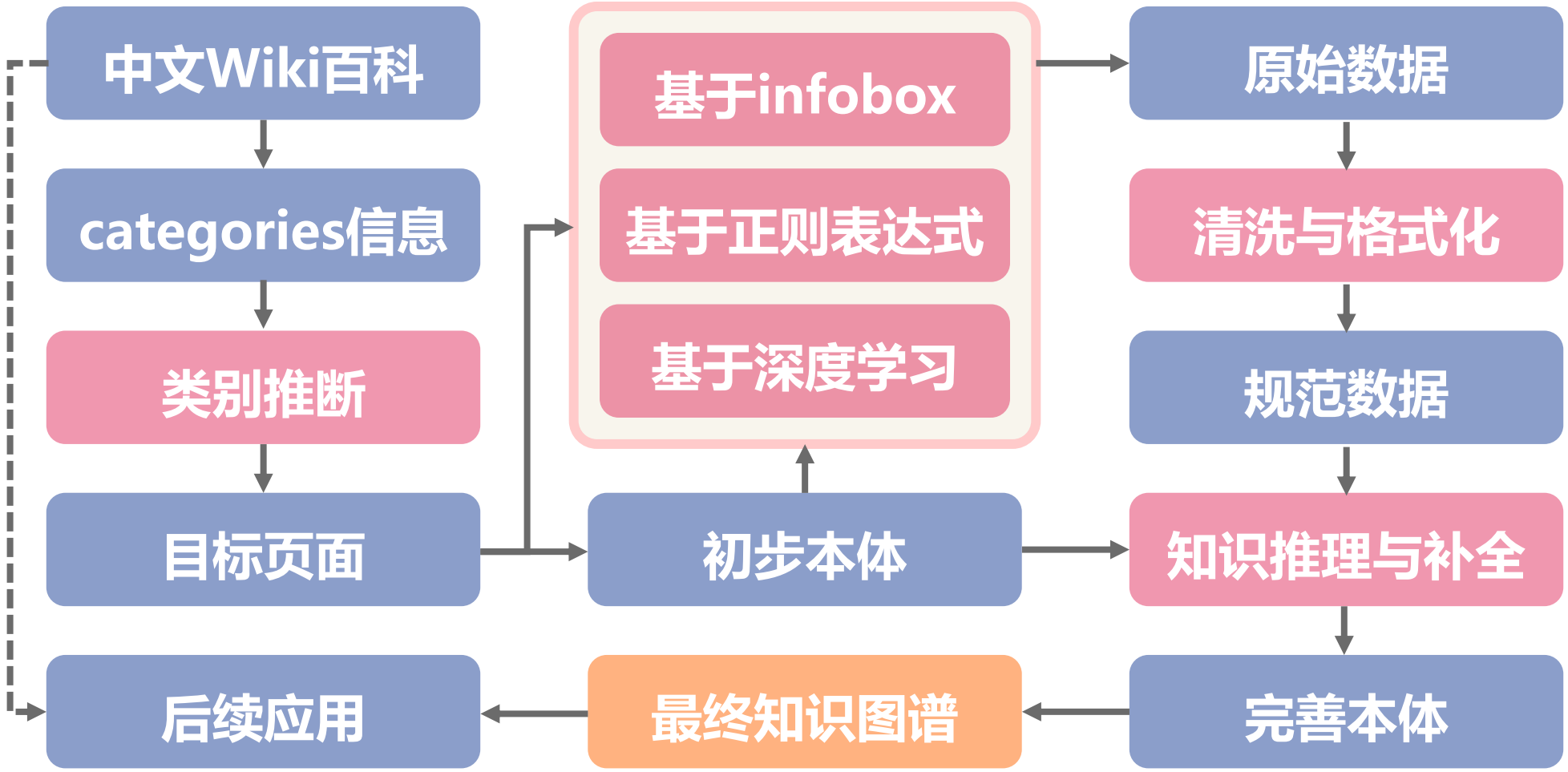
第15组 答辩人：黄一凡 指导老师：吴天星

目录

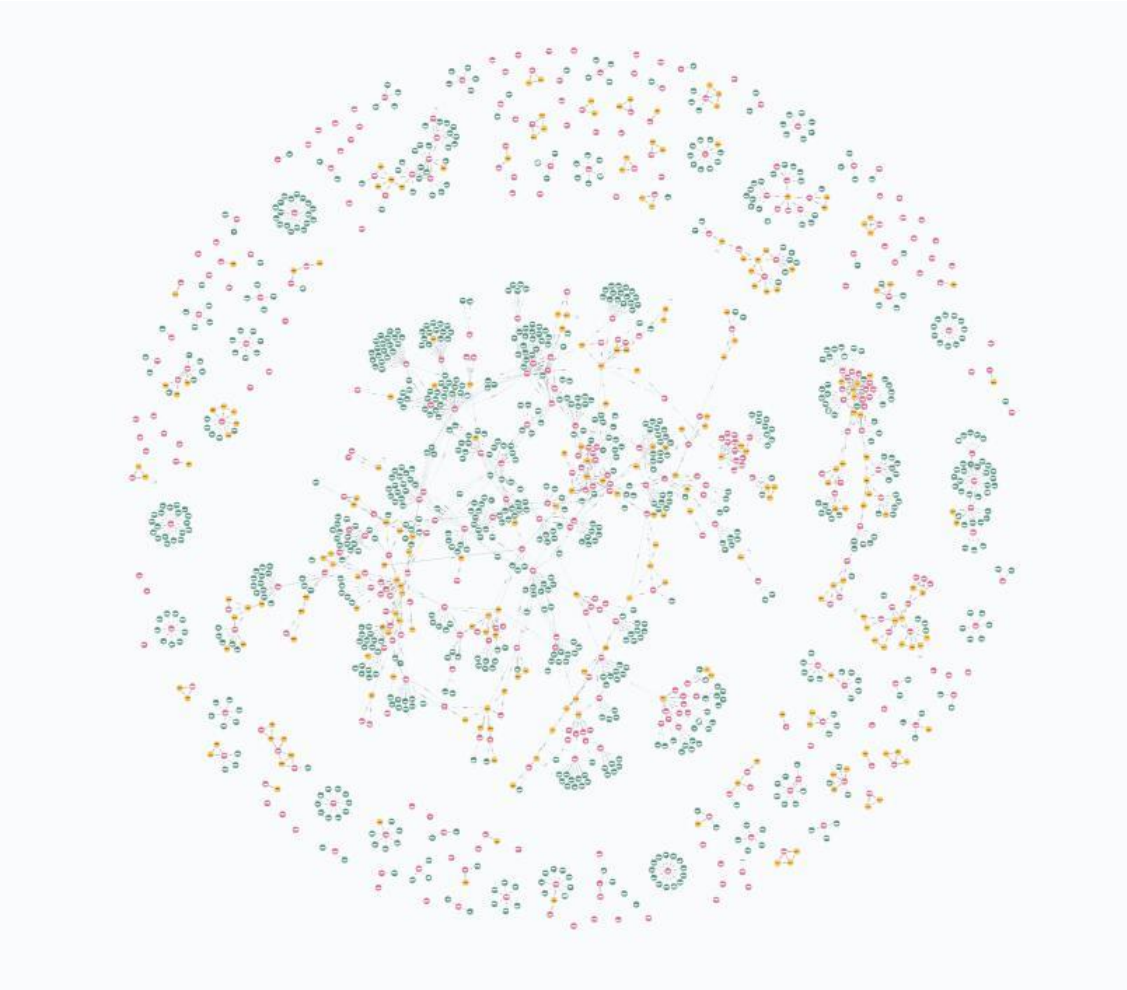
第一部分

整体概况

整体框架


中文Wiki百科
categories信息
类别推断
目标页面
后续应用
基于infobox
基于正则表达式
基于深度学习
初步本体
最终知识图谱
原始数据
清洗与格式化
规范数据
知识推理与补全
完善本体

最终知识图谱



*该可视化为部分知识图谱（1500个实例）

本图谱包含：

- 153802个三元组
- 50622个实例（包含大量地址、亲属和电影实例，人物示例为13252个）

第二部分

实现细节

类别推断

基于每个页面的categories信息，确定其类别（演员、导演、编剧）

- 相较于infobox信息，categories信息更加全面、细致
- 其全面体现在有部分页面不存在infobox
- 其细致体现在infobox中的类别只有笼统的“艺人”，无法判断其具体身份（歌手or演员？）

建立推断规则R

对于页面p的categories集合$C_p = \{c_1, c_2, \dots, c_n\}$与类别$i$的规则$r_i \in R$，存在这样一个$c_k \in C_p$：在$c_k$中包含$r_i$中所有的关键词且不包含$r_i$中任意一个违禁词，那么我们认为页面$p$属于类别$i$。

类别推断

推断规则R

类别	关键词	违禁词
电影演员	电影；演员	协会；处女作；奖
电影导演	电影；导演	协会；处女作；奖；导演电影
电影编剧	电影；编剧	协会；处女作；奖；编剧电影

[[Category:英皇娱乐艺人]]
[[Category:香港电影女演员]]
[[Category:圣芳济各书院校友]]

演员

[[Category:20世纪电影]]
[[Category:王家卫导演电影]]
[[Category:电影演员处女作]]

非演员

接着，由推断出的类别可以进一步确定中文维基中哪些页面是我们需要的

初步本体构建

使用自底向上的方法构建本体

1. 抽取某一类（演员、导演、编剧）目标页面所有infobox中的属性
2. 对infobox和消歧后的各属性分别计数，记为$count(infobox)$和$count(p_k)$
3. 对于某个属性i，若$\frac{count(p_i)}{count(infobox)} > \frac{1}{3}$，我们就认为属性$i$是该类的属性

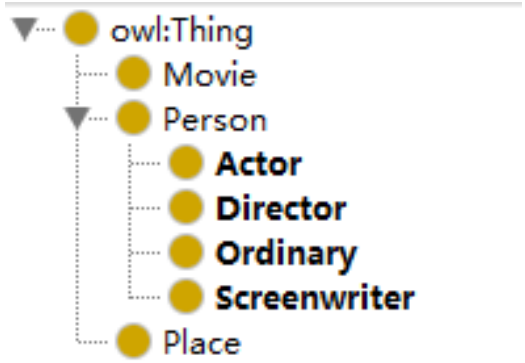
属性	count	count/infobox
Infobox	10471	1.00
IMDb	8576	0.82
奖项	5462	0.52
国籍	6303	0.60
墓地	1499	0.14
……	……	……

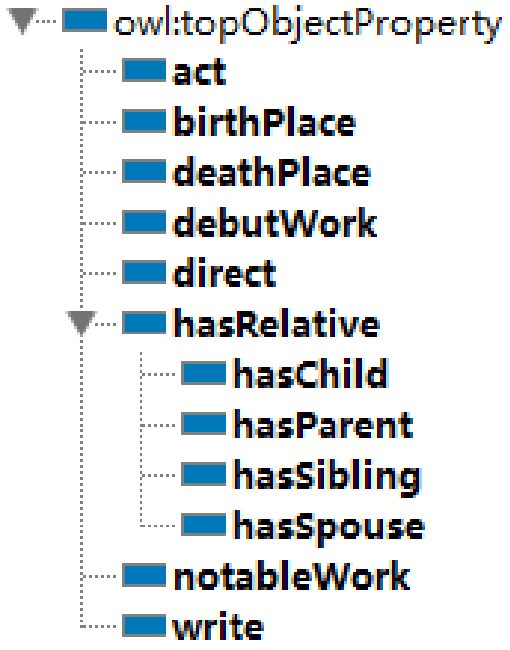
*以演员为例：

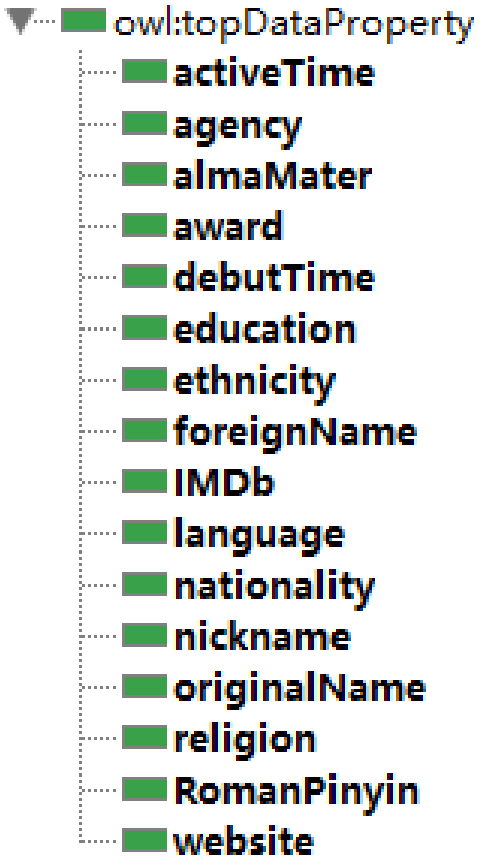
为演员类的属性

不为演员类的属性

初步本体构建

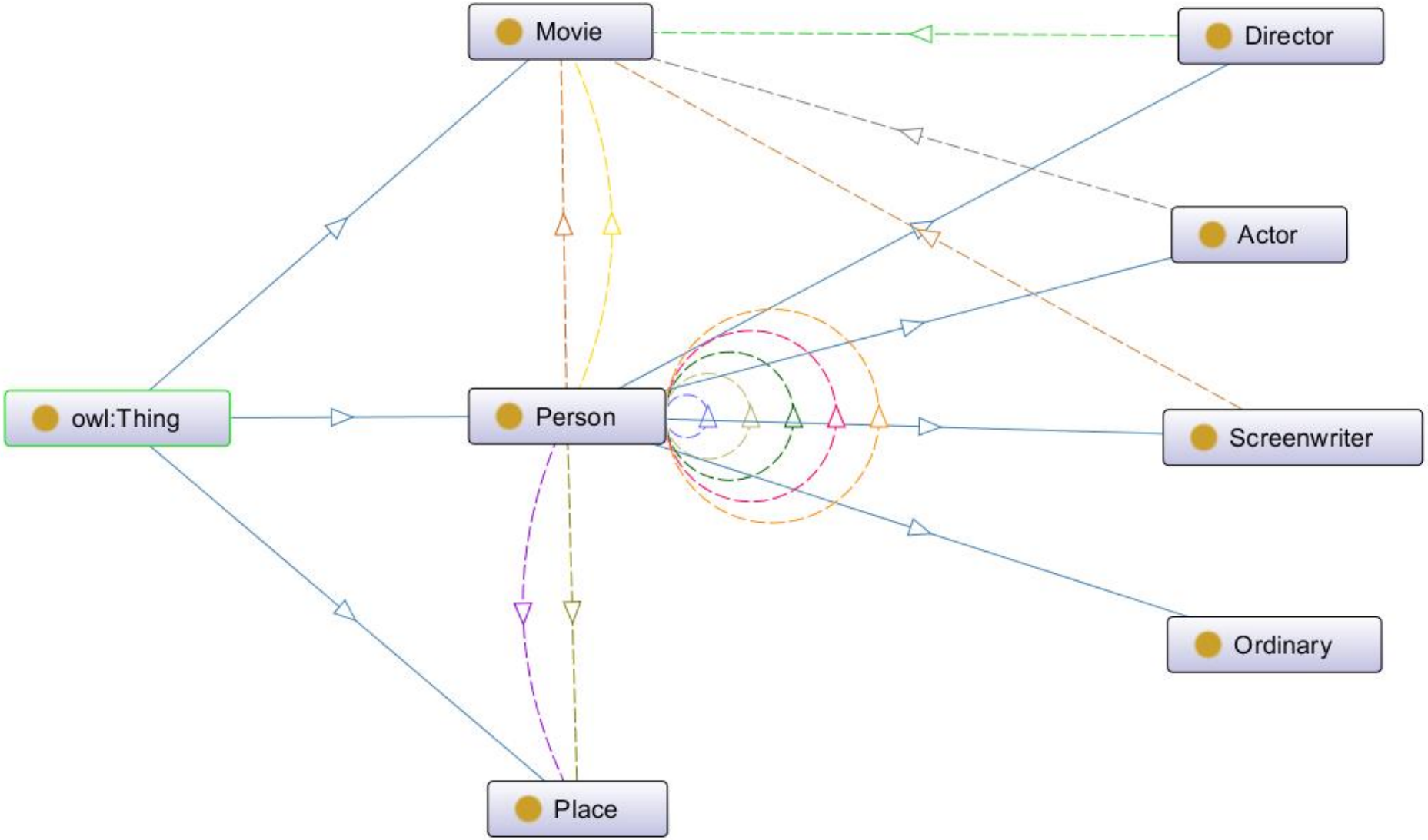






共有7个类别，28个属性（其中Object Property*12，Data Property*16）

初步本体构建


Movie
Director
Actor
owl:Thing
Person
Screenwriter
Ordinary
Place

事实抽取

基于infobox的事实抽取

- 根据部分页面中存在的infobox直接进行事实抽取

女艺人	
英文名	Faye Wong（1991年至今） Shirley Wong（1989年 - 1991年）
昵称	阿菲·菲姐·天后·Pop Diva [1] [2]
别名	夏林（15岁之前使用） 王靖雯（1989年 - 1996年使用）
国籍	中国（香港）
出生	1969年8月8日（51岁） 中华人民共和国北京市北京协和医院[3]
职业	歌手·演员·音乐制作人·词曲作家·慈善家
语言	国语·粤语·英语·日语·梵语
母校	北京市东直门中学
宗教信仰	藏传佛教 (噶举派)
配偶	窦唯（1996年结婚；1999年离婚） 李亚鹏（2005年结婚；2013年离婚）

```
{{艺人
| 姓名 = 王菲
| 类型 = 女艺人
| 图片 = Faye 2011 Hong Kong cropped.jpg
| 图片尺寸 =
| 图片简介 = 王菲在演唱会上
| 英文名 = Faye Wong（1991年至今）&lt;br /&gt;Shirley Wong（198
| 暱称 = 阿菲·菲姐·天后·Pop Diva &lt;ref&gt;{{Cite web |url=htt
| 别名 = 夏林（15岁之前使用）&lt;br&gt;王靖雯（1989年－1996年使用
| 国籍 = {{CNHK}}
| 出生日期 = {{Birth_date_and_age|1969|8|8}}
| 出生地点 = {{PRC}}[[北京市]][[北京协和医院]]&lt;ref&gt;{{Cite
| 职业 = [[歌手]] ·[[演员]] ·[[音乐制作人]] ·[[词曲作家]] ·[[慈善
| 语言 = [[国语]]·[[粤语]]·[[英语]]·[[日语]]·[[梵语]]
| 宗教信仰 = [[藏传佛教]] ([[噶举派]])
| 配偶 = {{marriage|[[窦唯]]|1996|1999|end=div}}&lt;br&gt;{{ma
```

事实

缺点：存在约$\frac{1}{10}$的页面没有infobox，且infobox中信息不全

事实抽取

基于正则表达式的事实抽取

- 使用正则表达式对网页中的text数据进行事实抽取
- 观察本体中各属性的特点
- 根据特点编写相应的正则表达式

属性	正则表达式
hasChild	(?<=(女儿|儿子|孩子|小孩)).*?(?=(出生|诞生|降临|是))
almaMater	(?<=就读于)[\u4e00-\u9fa5]+?(学校?)；(?<=(于|在))[\u4e00-\u9fa5]+?(学校?)(?=就读)
nickname	(?<=绰号|又叫|也叫|绰号叫|绰号是)[\u4e00-\u9fa5]+
……	……

缺点：抽取出的数据较为杂乱，且经常会抽出一整个句子

事实抽取

基于深度学习的事实抽取

- 使用工具jiagu对网页中的text数据进行三元组关系预测
- 其模型训练使用https://github.com/ownthink/KnowledgeGraphData作为训练数据

原文本：阮玲玉（），原名阮凤根、训名学名阮玉英，祖籍广东省香山县，生于上海县上海，中国无声电影默片时代演员。她是1930年代中国影坛最突出的明星之一，其优秀的演技与于24岁时自杀一事使之成为中国电影的一个时代象征。

提取出的三元组：[阮玲玉, 祖籍, 广东省香山县]
[阮玲玉, 出生地, 上海县]
[阮玲玉, 国籍, 中国]

数据清洗与格式化

动机：抽取出的事实较为杂乱，且许多信息隐藏于字符串中，难以利用

基本方法：

- 按顿号或者逗号将输入字符串切分为多个字串
- 清除无效字符以及链接
- 删除无意义词（如“电影”、“电视剧”）
- 删除括号
- 使用正则表达式对目标内容进行匹配（如“《》”、“奖”）
- 对数据进行格式化

数据清洗与格式化

1、日期数据格式化（debutTime;activeTime）

'1997年~2002年，2010年至今' → ['1997年-2002年' ， '2010年-2021年']

方法：提取数字并将至今替换为2021

2、姓名数据格式化（nickname;originalName;foreignName）

'小美、叶小美、青儿、牙签黄瓜' → ['小美', '叶小美', '青儿', '牙签黄瓜']

方法：用顿号将输入字符串切分为多个字串

3、作品数据格式化（notableWork,debutWork）

'月满西楼（1968年）\n 庭院深深（1971年）' → [' 《月满西楼》', '《庭院深深》']

方法：删除无效字符及括号，并加上书名号进行格式化

4、奖项数据格式化（award）

'土星奖最佳男主角1996年《杀出个黎明》 国家评论协会最佳男主角2007年《全面反击》 →

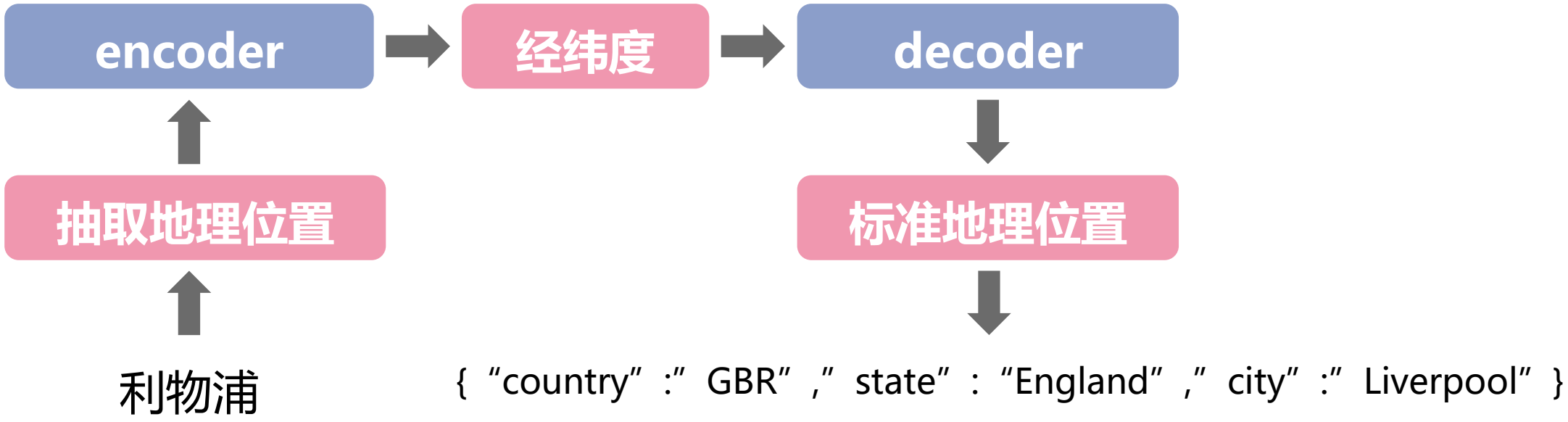
['土星奖最佳男主角', '国家评论协会最佳男主角']

方法：使用正则表达式提取目标内容“奖”、“最佳”等

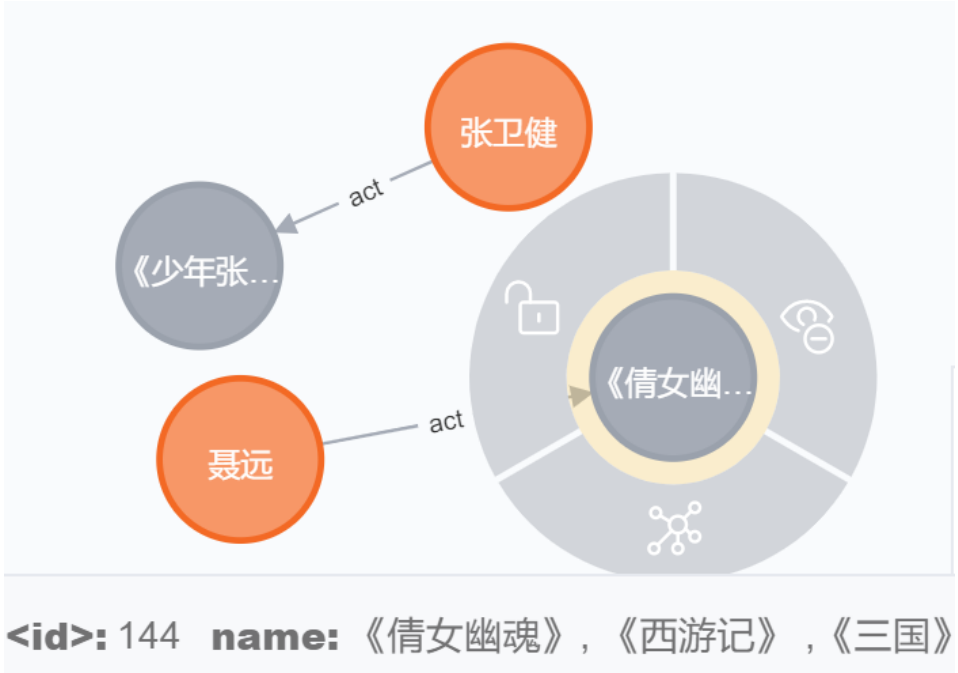
数据清洗与格式化

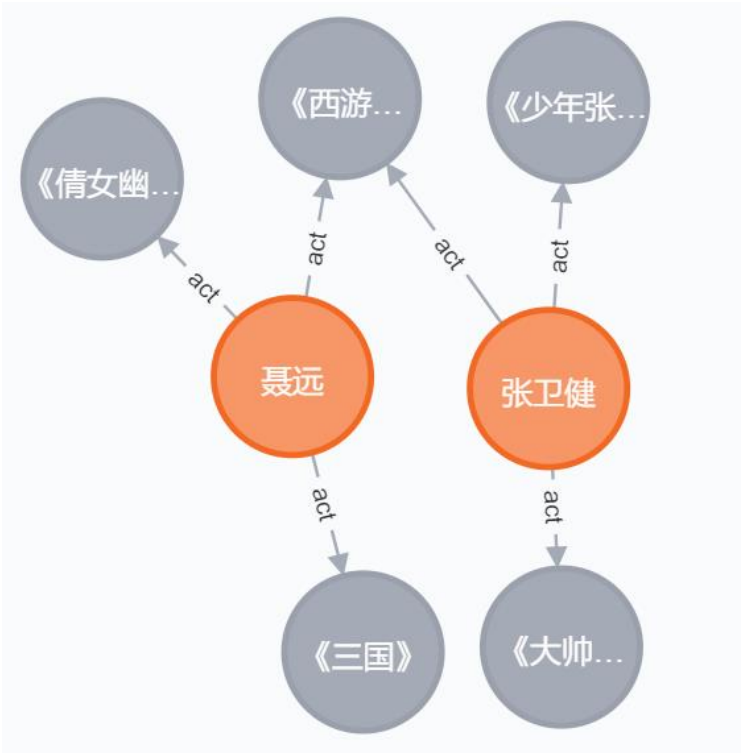
5、地理数据格式化（birthPlace;deathPlace;nationality）

- 使用工具geocoder
- 利用经纬度将抽取出的地理位置转化为标准格式
- 以字典形式返回，便于后续的应用（如查询、推理）



数据清洗与格式化



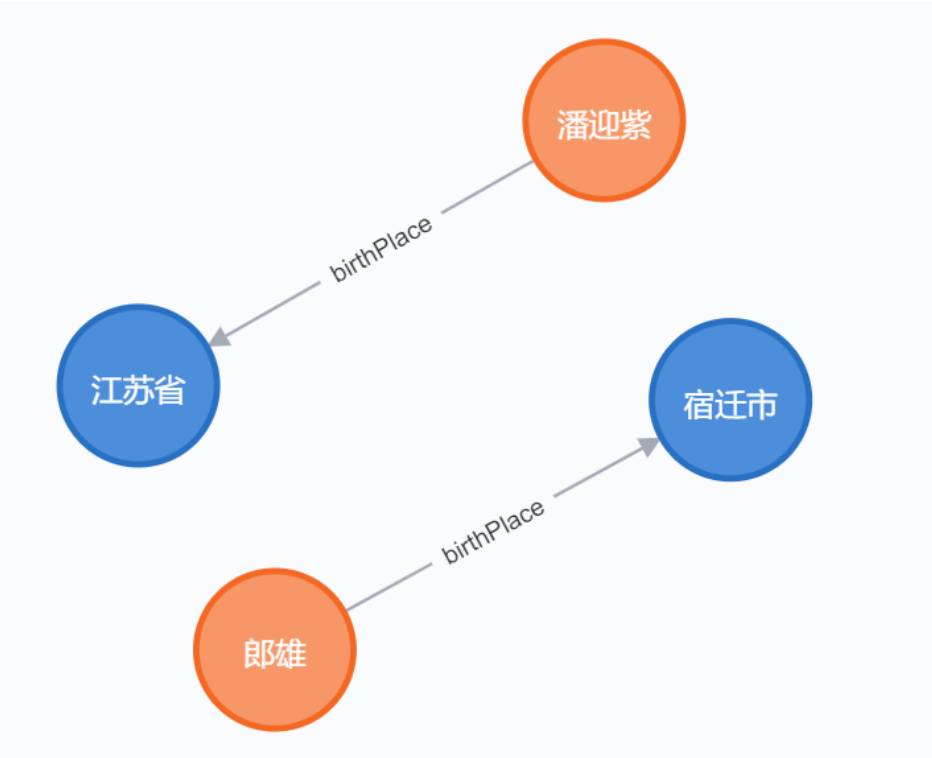


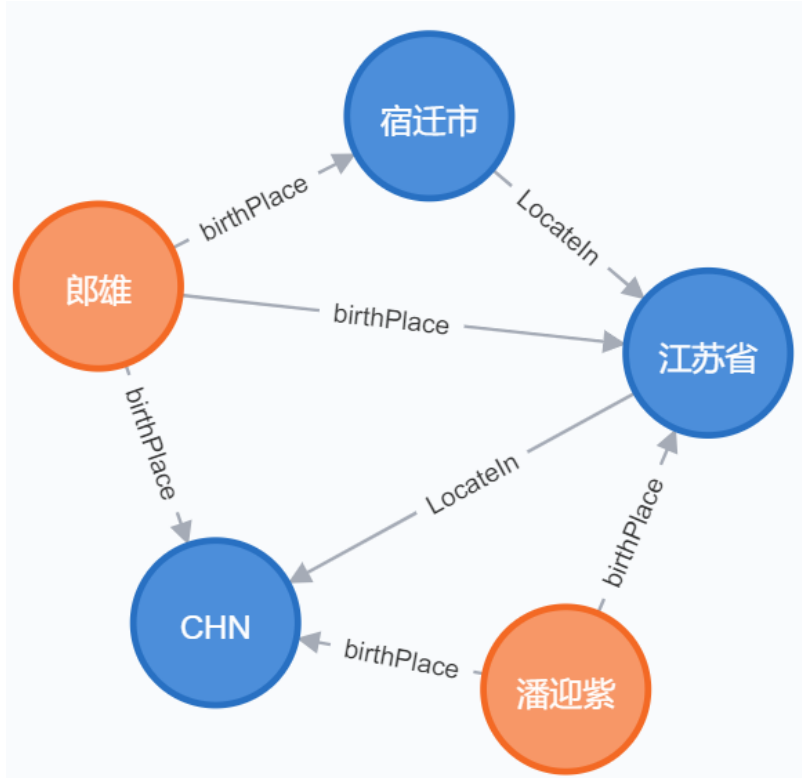
清洗格式化前的出演数据

清洗格式化后的出演数据

- 格式化前出演的电影名称杂糅在一个字符串内
- 格式化后将每一部电影分开，可以发现演员间的合作关系

数据清洗与格式化





清洗格式化前的地理数据

清洗格式化后的地理数据

- 格式化前的地理数据只是简单的字符串，包含信息较少
- 格式化后的地理数据是结构化的字典，可以包含更多的信息

知识推理与补全

动机：经过事实抽取、数据的格式化和清洗，我们可以根据已有的图谱建立新的关系，从而进一步提高知识图谱的完备性，并为今后可能的应用奠定基础

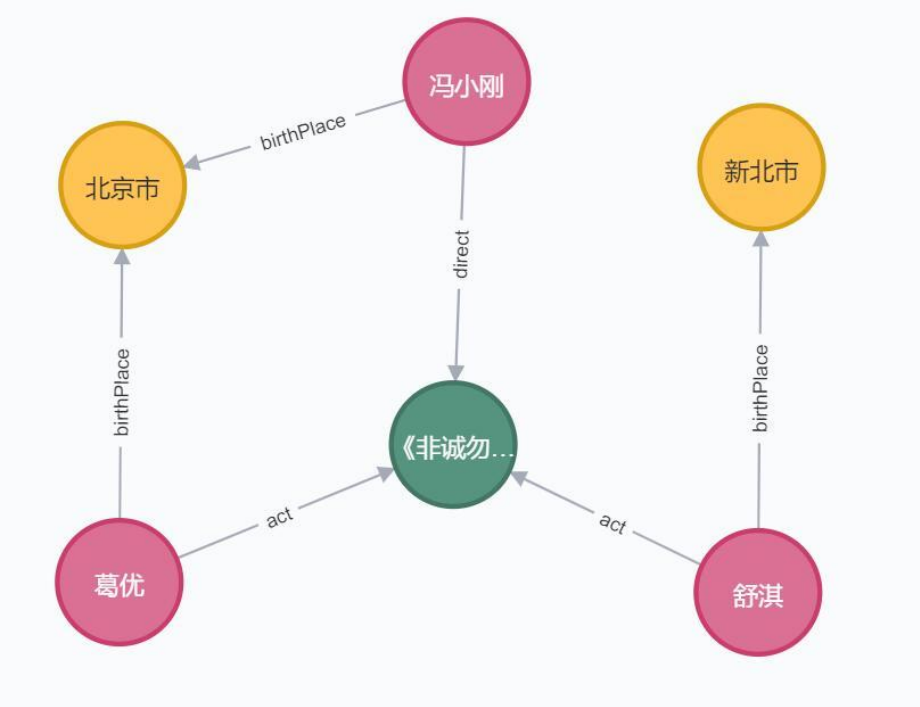
补全内容：

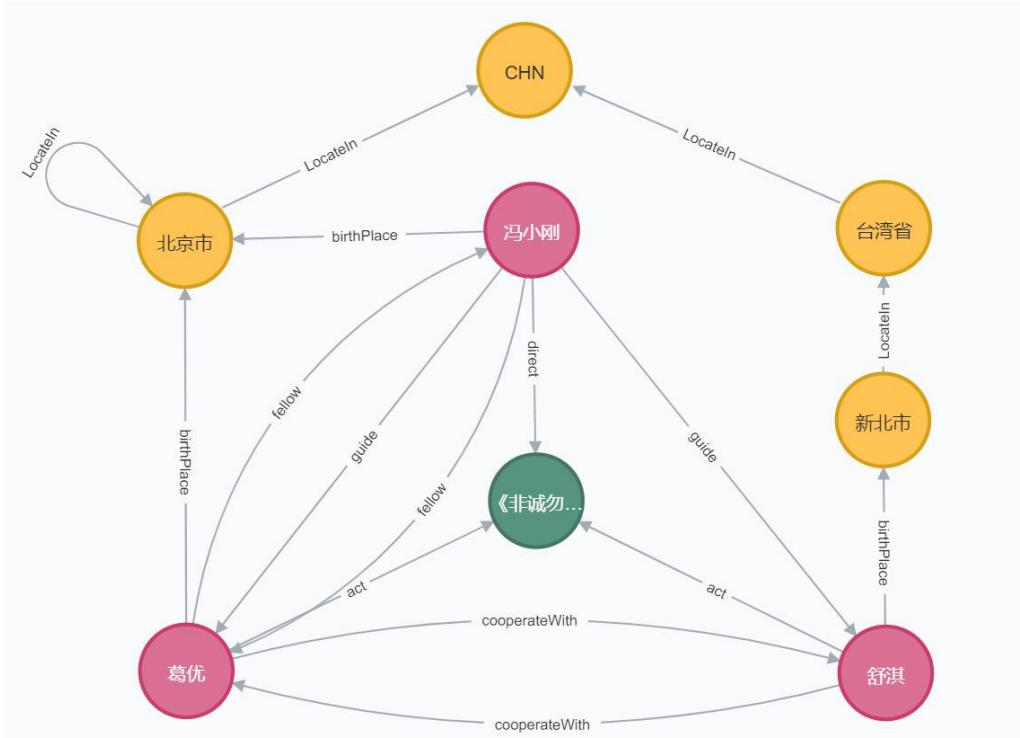
- 若两个演员参演了同一部电影，则在两个演员之间添加’ cooperateWith’ 关系。
- 若两个人出生地相同，则在两个人之间添加’ fellow’ 关系。
- 若导演和演员共同参与了一部电影，那么这位导演与演员之间创建’ guide’ 关系。

Cypher示例：

```
MATCH (a1:actor),(a2:actor), (m:Movie)
WHERE (a1)-[:act]-(m) and (a2)-[:act]-(m) and a1<>a2
CREATE (a1)-[r:cooperateWith] -> (a2)
```

知识推理与补全



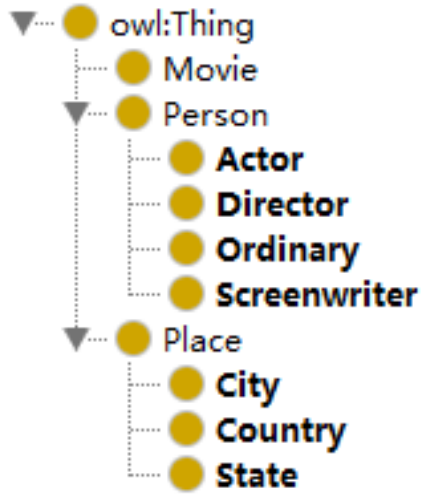


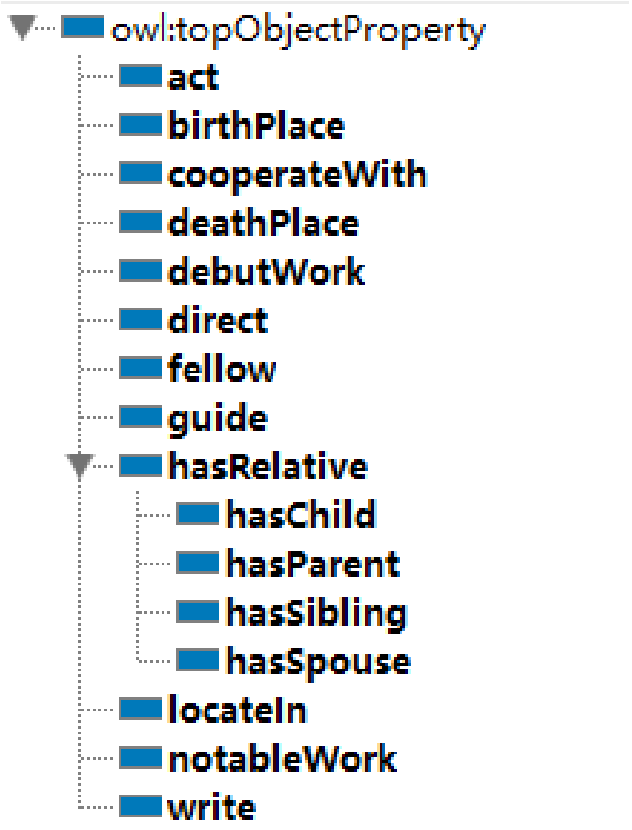
知识补全前

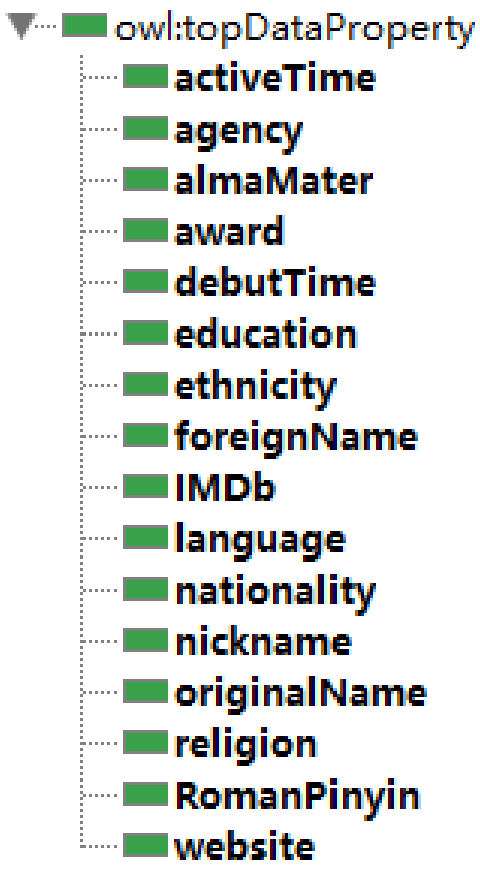
知识补全后

- 知识补全前人物之间没有关联，无法显示人物之间隐含的关系。
- 知识补全后可以找出人物的老乡、合作伙伴和指导关系。

完善本体构建

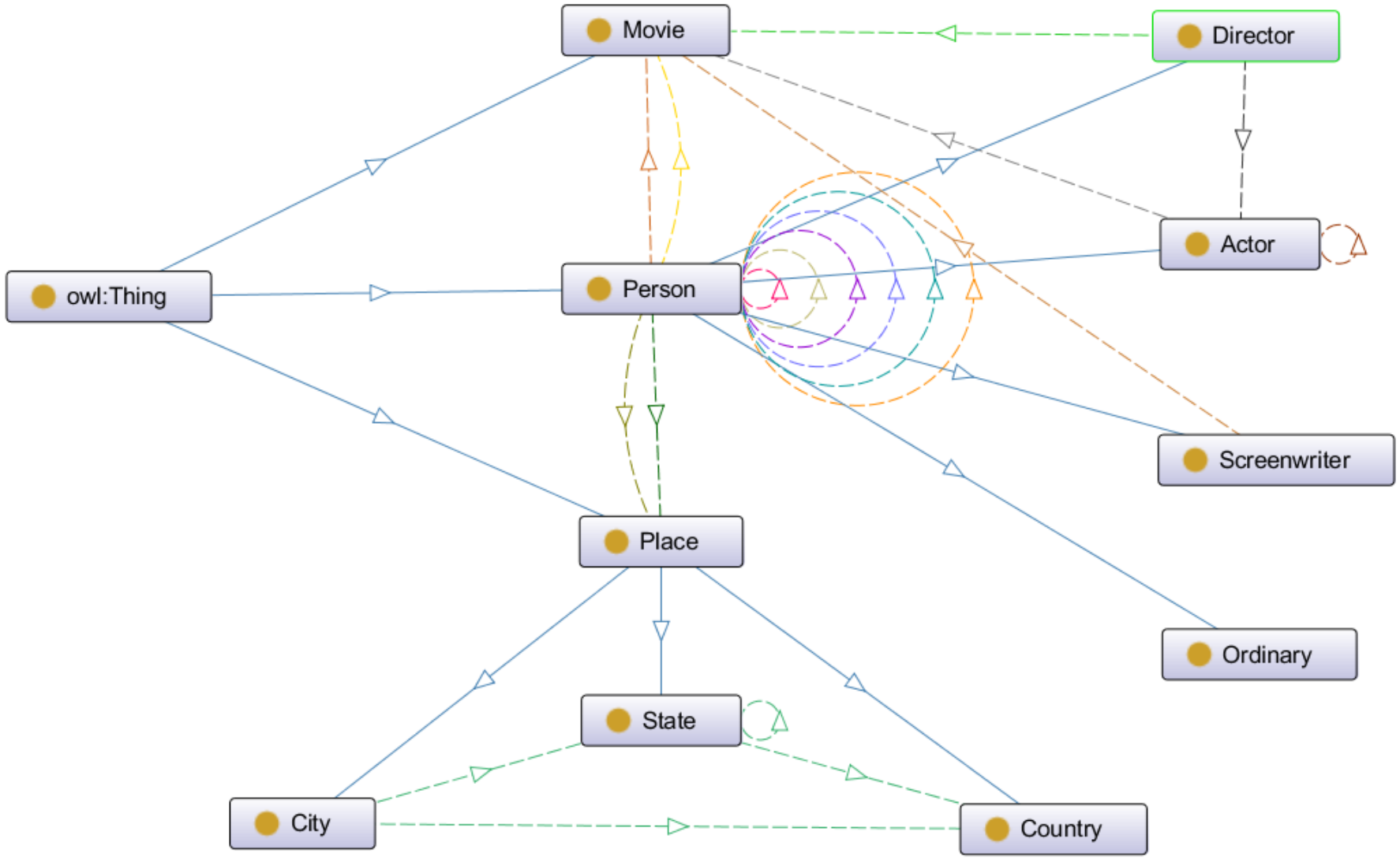






完善后共有10个类别，32个属性（其中Object Property*16，Data Property*16）

完善本体构建


Movie
Director
owl:Thing
Person
Actor
Screenwriter
Place
Ordinary
State
City
Country

第三部分

团队分工

分工

	解析数据	类别推断	本体构建	事实抽取	数据清洗	知识补全	可视化
黄一凡							
曹思辰							
唐云龙							
徐浩卿							
张妍							
谈笑							

感谢各位评判指导

第15组 答辩人：黄一凡 指导老师：吴天星