《哈利·波特》知识图谱

夏鹏、朱烨、蒋逸凡

指导老师：李直旭

苏州大学计算机科学与技术学院

目录

1. 摘要

针对于构建主题型知识图谱，哈利·波特魔法世界人物关系层次适中，较为符合作业要求，且哈利·波特广为人知，方便爬取数据。因此，我们选择哈利·波特作为图谱主体。

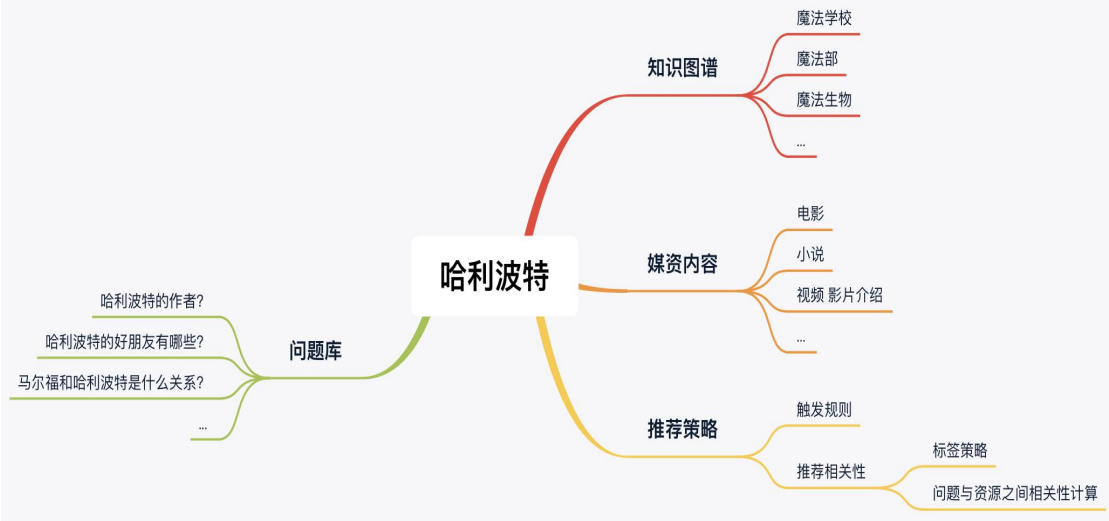
我们的需求是将哈利·波特魔法世界人物按照魔法学校、魔法部、魔法生物、其他进行分类，并尝试使用自动化的方法来实现知识图谱的构建。

2. 分工

- 夏鹏：实体的属性载入，问答库的爬取、清洗与整理
- 蒋逸凡：多媒体数据的爬取、清洗、梳理，论文的部分完善
- 朱烨：原始实体数据的获取与补全，protege 的初步界面化

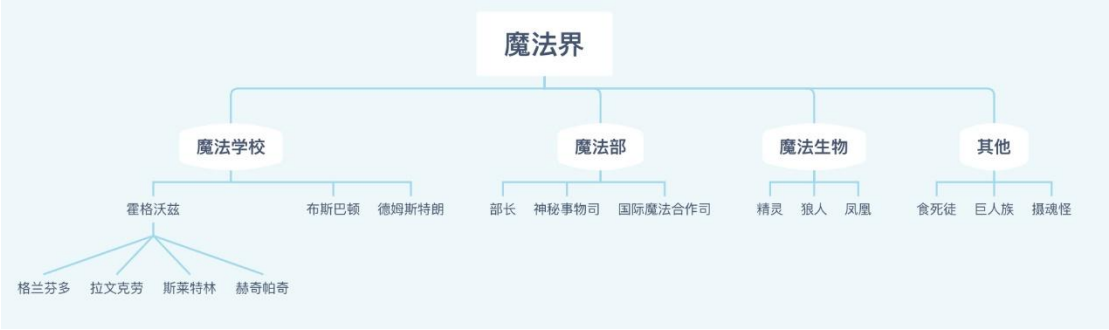
3. 图谱概况

3.1 基本结构

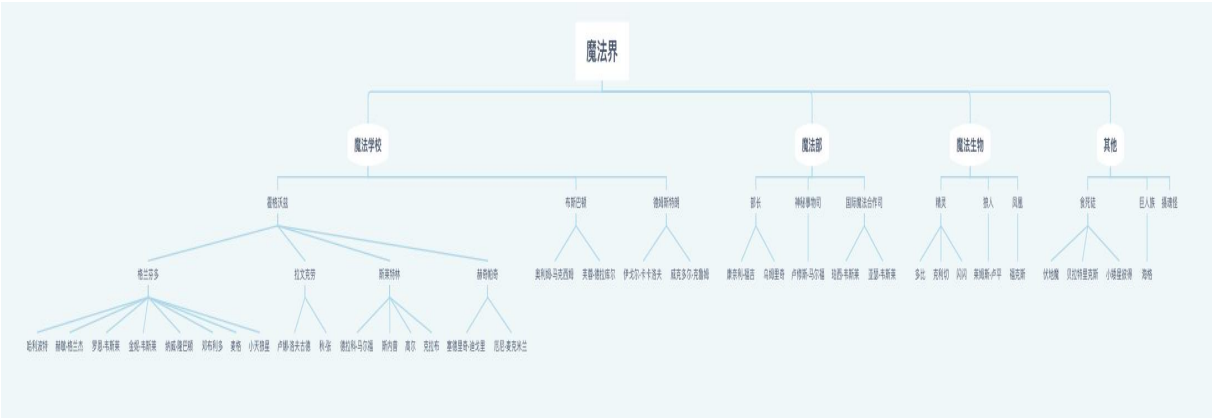


3.1 图：基本结构

3.2 概念树



3.2 图 1：概念层



3.2 图 2：实体层

4. 数据收集与处理

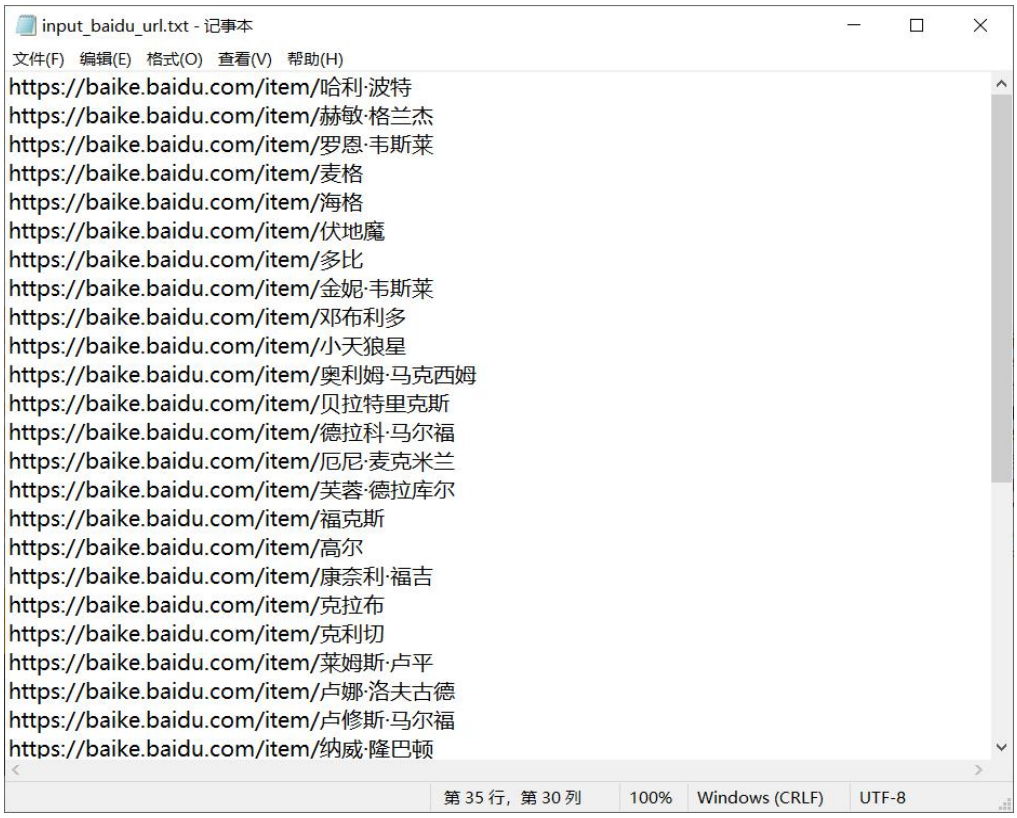
4.1 数据准备与爬取

我们收集了哈利·波特中出现的主要人物，将他们作为实体。把人物按照学院派别、魔法部、魔法生物、其他生物形成关系树，并梳理人物间的关系。在将所有实体组建成概念树以后，通过百度百科爬取各个人物的基本属性，在进行数据清洗后我们最终选择性别、生日、外貌、性格特点、魔杖、饰演、别名这七个标签。作为每个人物的基本属性，并对所有实体的基本属性内容进行完善。

4.2 多媒体数据

人物图片数据爬取：根据所有人物名字列表，根据维基百科网址的生成格式，生成所有人物百科网页网址列表。接着解析网页并获取图片链接，在爬取每个人物的图片后判断是否为空，如果为空则代表第一次爬虫未能爬到图片，该人物名字加入第二轮爬虫的列表。

人物视频数据爬取：大概步骤与爬取图片步骤相似，爬取 mp4 文件，但由于受到内存大小限制，只选取了 5 个主要重点人物的百度百科人物介绍视频。



4.2 图 1：百度百科人物网址列表

4.3 QA 问答库

爬取了一些相关的垂域网站，选取了大概 100 左右的问题，首先筛去一些字数过多的问题，然后筛去很难找到 tag 的问题，这样我们就可以为每个问题选择合适的 tag，挂载实体层属性，并生成 json 文件。

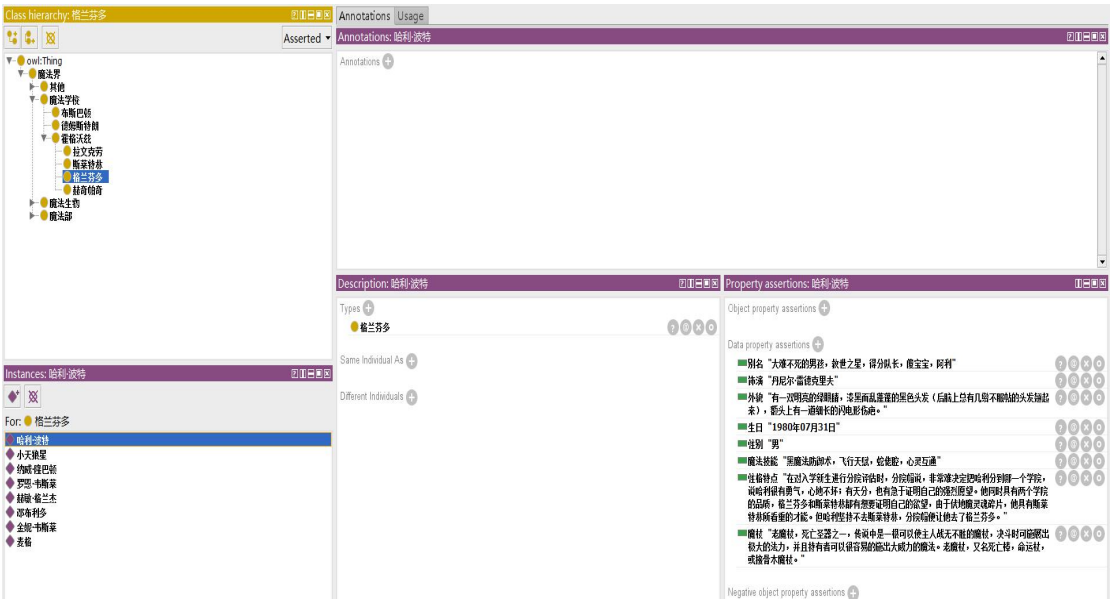
5. 图谱构建

我们将介绍图谱构建的过程，包括对初始数据处理的一些步骤，问答库和媒体库的爬取及挂载等。

5.1 实体属性载入

将概念层所属关系、实体属性挂载，生成 owl 文件，导入

Protégé 效果如下



5.1 图 1：导入 Protégé 效果图

5.1 图 2：owl 文件

然后我们输出三元组，自动挂载属性三元组以及关系三元组。

5.2 生成 json 文件

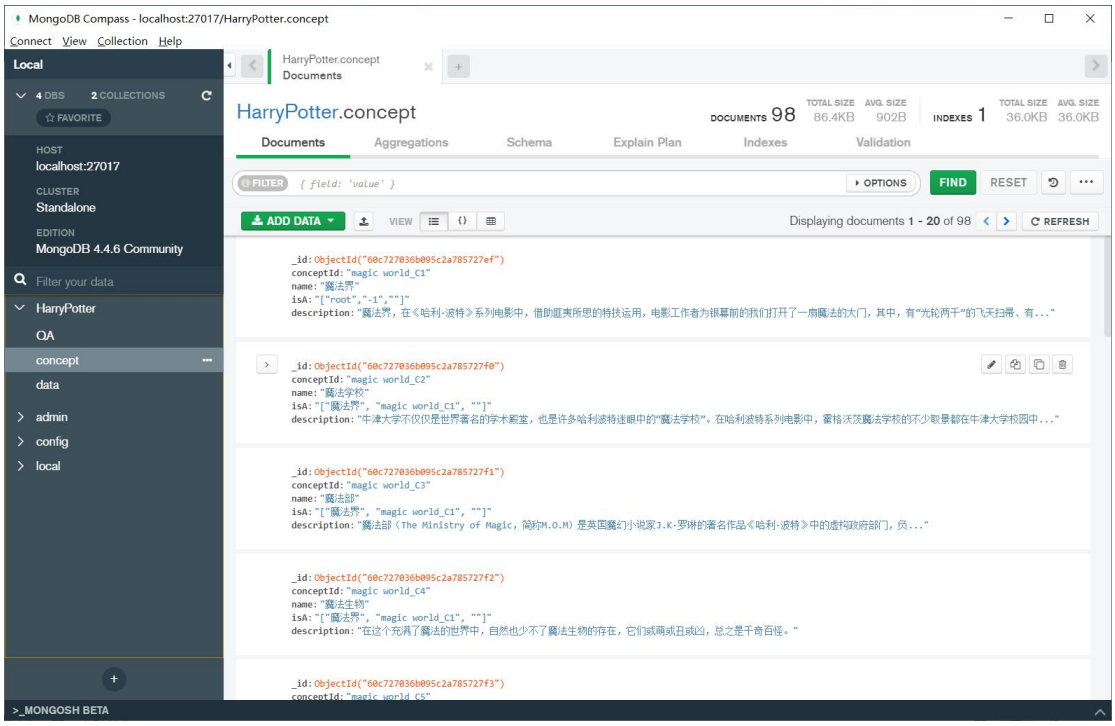
我们使用 MongoDB 来进行数据库的存储和管理，导入 csv 文

件，使用 compass 可视化工具进行 json 文件导出。效果图如下。

```python
# -*- coding: utf-8 -*-
import json

f=open("concept.csv","r",encoding='utf-8')
ls=[]
for line in f:
        line = line.replace("\n", "")
        ls.append(line.split(","))
f.close()
print(ls[1:])
fw=open("concept.json","w",encoding='utf-8')
for i in range(1,len(ls)):
    ls[i]=dict(zip(ls[0],ls[i]))
a = json.dumps(ls[1:],sort_keys=True,indent=4,ensure_ascii=False)
print(a)
fw.write(a)
fw.close()
```

5.2 图 1：csv 文件转 json 文件 code



5.2 图 2：json 导入数据库效果

6. 心得体会

知识图谱描述了知识资源及其载体之间的相互联系，利用知识图谱可以找到最想要的信息，最全面的概要。

本次知识图谱大作业，我们以哈利波特为主题，构建了一个简易的主题型图谱，将哈利波特魔法世界进行人物分类，将人物作为实体，构建人物之间关系，外貌，性格特点，魔杖，生日等属性关系。

但美中不足，自动化程度还不够高。另外由于时间问题，未能将推荐系统结合其中，否则结合 KBQA 可以尝试实现一个简易的问答系统。

7. 结语

此次大作业，我们针对哈利波特这一主题在网上进行数据的搜集、筛选与清洗，并结合在知识表示课上所学到的知识，制作了一个哈利波特主题型知识图谱。附上了我们的多媒体资源，以及我们制作过程中的临时文件，以及最终的 json 文件。总体来说，作业的完成度较高。美中不足是较多部分是我们自己手动操作，而非自动化处理，希望能在以后的大作业或者实践过程中改善这点。

通过此次的大作业实践，我们深刻的感受到了实践的重要性，有些看起来很简单的工作做起来却有那么多的不足和改善的空间。这更让我们感受到知识图谱的厉害。他对对于

问题的处理、智能的交互有着如此大的作用，这将引领着我们在未来的学习过程中发挥不可替代的作用。