

2020-2021 学年

知识表示课程设计报告

任课教师：吴天星

院 系 _计算机科学与工程学院_

专 业 ____人工智能_____

组 别 ____第四组_____

知识图谱实验报告

第4小组

09118219 王一名, 09118228 沈飞鸿, 09118229 张博宇

09118230 鲁瀚洋, 09118236 廖滔, 09118241 陈嘉源

目录

一、效果介绍和展示	2
二、实现模块与算法原理	2
2.1 本体构建	2
2.2 事实抽取	3
2.2.1 问题: URL 定位	3
2.2.2 信息抽取	4
2.2.3 人物图片爬取	5
2.2.4 三元组关系	5
2.3 类别推断	6
2.3.1 Type Inference from Infoboxes	6
2.3.2 Type Inference from Text	7
2.4 知识融合	9
2.4.1 数据预处理	9
2.4.2 数据特征分析	10
2.4.3 基于半监督的规则学习算法	11
2.5 问答系统	14
2.5.1 分词模型	15
2.5.2 语义解析	15
2.6 推荐系统	16
2.6.1 数据解析	16
2.6.2 SOTA 分词模型	16
2.6.3 word2vec	17
2.6.4 模型分析	18
三、数据库可视化	19
四、心得体会	23
参考文献	25

一、效果介绍和展示

2019 年，电影《复仇者联盟 4：终局之战》以 27.97 亿美元的票房击败了詹姆斯·卡梅隆的《阿凡达》，成功夺得了影史票房 NO.1 的宝座。该影片取得这样的成绩不仅因为制作质量，更大的原因是背后有漫威这个 IP 的加成。自 2008 年起，漫威电影宇宙计划依次将钢铁侠、绿巨人、雷神托尔、美国队长等超级英雄搬上大荧幕，这些作品都获得了极高的人气，深受全球各国年轻人的喜爱。不过，漫威公司创造的超级英雄角色多达 5000 多个，基于已经上映的电影以及一些漫画很难对其架空世界以及超级英雄基本信息、人物关系等有一个清晰、结构化的了解。

本项目构建的漫威人物知识图谱基于美国漫威漫画公司制作的一系列漫画和电影组成的架空世界和共同世界。基于我们构建的知识图谱，用户可以结构化的查询人物信息，了解角色间的关系以及漫威超级英雄全貌，更深入地了解漫威宇宙。

本次课程设计，小组成员共同完成了面向百科的中文人物领域知识图片的构建，所选人物范围为漫威人物，包括复仇者联盟、神盾局特工与反派等角色类别。最终通过 neo4j 数据库完成了整体框架的可视化。图谱数据基于百度百科、维基百科以及漫威人物数据网页

实验完成的内容除了必做内容以外，还包括了选做内容的 Knowledge Fusion：将百度百科与维基百科的数据进行了融合。另外，在知识问答时，我们选取了中文分词的 SOTA 模型来对问题进行关键词划分。复现模型为来自 NeurIPS 的 N-LTP 和来自 ACL 的 WMSeg。

二、实现模块与算法原理

2.1 本体构建

本体是用于描述一个领域的术语集合，其组织结构是层次结构化的，可以作为一个知识库的骨架和基础。

也可以说，本体定义了组成「主题领域」的词汇表的「基本术语」及其「关系」，以及结合这些术语和关系来定义词汇表外延的「规则」。

1. 领域：「漫威人物」；
2. 基本术语：「超级英雄」、「反派」等等概念
3. 关系：超级英雄之间为朋友关系，反派与超级英雄之间为敌对关系，另外也存在一些情人/爱人关系。
4. 类别：复仇者联盟、神盾局特工、银河护卫队、超级反派、其他
5. 属性：中文/英文名、性别、登场作品、所属团队等

2.2 事实抽取

百科数据中，有许多关于漫威人物的信息，其中有不少冗余信息，如人物冗长的经历与生平等。以百度百科为例，有用的信息为人物之间的关系、阵营，人物的基本信息以及能力值的表格。



图 1 人物基本信息

能力数值		
智力	3	力量
速度	2	敏捷
耐力	3	战斗技巧
说明：MARVEL漫威官方能力参考		
属性	备注	分析
智力	思考和处理信息的能力	1. 迅速解码、2. 正常、3. 学识、4. 超天赋、5. 天才、6. 超级天才、7. 无所不知
力量	举起沉重物体的能力	1. 极弱（不能举起日常用品所重量的物体）、2. 正常（能举起与自身体重相等的物体）、3. 常人级别（能举起体重是常人两倍的物体）、4. 超级常人（800kg-25t）、5. 超级巨人（25t-75t）、6. 超级巨人（75t-100t）、7. 极限力量（超过100t）
速度	在地球上移动的能力	1. 正常以下、2. 正常、3. 超级常人（能不善于时速700英里）、4. 高速（1马赫）、5. 超高速（3马赫）、6. 光速（每秒186000英里）、7. 极速太真（超光速）
耐力	抗击或吸收伤害的能力	1. 极弱的、2. 正常、3. 超级的、4. 司母的、5. 特别强的、6. 超出常人、7. 近乎不死的
耐力	向线程施加能量的能力	1. 极弱的能量、2. 通过接触就能能量、3. 在短时间内释放耗尽的能量、4. 在短时间内释放中等的能量、5. 在短时间内释放消耗的能量的一半能量、6. 能同时释放多种能量形式、7. 能在近乎无限的时间内释放各种能量形式
耐力	往手指出的强度和战斗技能	1. 不擅长、2. 一般的、3. 熟练过的、4. 有经验的战士、5. 精通一种格斗方式、6. 精通几种格斗方式、7. 精通全部格斗方式

图 2 人物能力数值表

2.2.1 问题：URL 定位

问题分析 由于给定一个人名，在百度百科中的网页并不是唯一的，存在其他与本次主题不相关的人物，因此仅仅根据人名定位网页的方法容易误选到别的人物。如图所示：



图 3 黑寡妇界面的多义项

问题解决 我们组首先确立总共所需要的漫威人物列表：利用网上爬取的方式，在漫威人物网站中爬取人物人名信息。接着利用**词条分类**的方法，观察各个词条与漫威之间的

关系，发现“漫威”二字本身不可用于唯一标识，但是“漫威旗下”或者“美国漫威漫画旗下”这些字眼却可以对漫威人物进行唯一性筛查。

所以我们依据人名进入百度百科的多义词界面，根据关键词”美国漫威漫画旗下”或”漫威旗下”进行正则化筛选，从而得到有用的 URL，将其保存于 Excel 表格之中，为之后的爬取工作提供正确的 URL 网址。

图 4 Excel 存储超级英雄和对应的 URL

2.2.2 信息抽取

Infobox 抽取 对于漫威人物的基本信息，我们首先利用百科的 Infobox 进行爬取。找到百科页面查看源网页得知，Infobox 的值以 basicInfo-item value 作为标签存储，因此代码如下：

```
key = soup.find_all(class_="basicInfo-item name")
val = soup.find_all(class_="basicInfo-item value")
```

这样一来，百度百科本身含有的人物基本信息就全部爬取下来了（如图 1 所示的基本信息）。

表格内容抽取 除了基本信息以外，还有图 2 所示的人物能力信息需要爬取。利用源网页信息中的标签可以知道，表格信息是存储在 `<table log-set-param="table_view" data-sort="sortDisabled">` 的标识里面，但用这样的方式直接爬取会爬取到所有的表格信息，从而带来了大量的无用信息。

因此我们制订了基于模版的筛选方法，只筛选符合模版智力-力量-速度-耐力-能量发射-战斗技能的表格，用该表格的表头信息进行字符串匹配以后成功得到了有用的能力值。

2.2.3 人物图片爬取

人物基本信息与关系三元组以后，我们将百科界面中漫威人物对应的图片爬取下来，存储于后端，在查询时显示出来，提供一个较好的可视化效果。

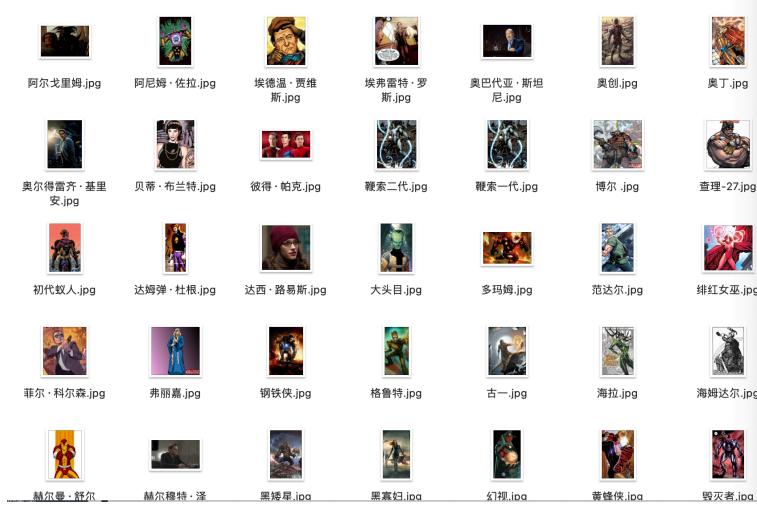


图 5 漫威人物图片存储

2.2.4 三元组关系

对于构建的知识图谱中人物关系三元组，主要数据源于外网某一漫威人物数据库，我们爬取其网站中存储的三元组 csv 文件，构建如（美国队长, 朋友, 钢铁侠）形式的人物关系三元组，但是由于该网站最后更新时间为 2018 年，且其人物关系是基于上映的漫威系列电影，导致该部分的人物关系存在时效滞后、内容不完整等问题。因此基于已有三元组关系，我们进一步通过 DeepKE 从爬取的百度百科中的角色经历中进行关系三元组抽取，将得到的三元组与原有的进行更正与补充。

DeepKE 是基于深度学习的开源中文关系抽取工具，我们对其中基于 CNN、Transformer、RNN 三个网络的关系抽取模型进行实验。在去年知识表示与推理课程中，漆桂林老师带领我们进行了知识抽取框架实践，根据实验结果我们发现对于是否给定头尾实体，三元组抽取的效果会有很大差别。因此在进行抽取时，我们将已有三元组的主语和宾语作为 head 和 tail。另外，在爬取的人物角色经历中，对于冗长的文本，我们使用正则表达式对主语和宾语都在某一句中出现的句子做关系抽取。

关系抽取数据集来源于苏州大学开源人物关系数据集。我们先对数据做预处理，筛除掉无效的字符标点，以及数字等冗余信息，并将实体对进行编码，这可以避免模型学习对实体词过分关注以至于导致 bias。我们采用了 one-hot 编码，将每个 token 映射为一个 id，构建词典存储所有的 token。最终可以把关系抽取出来，此处展示几个范例。将得到的关系与原本的关系进行对比，若两者存在不同，则进行人工判断，确认哪个关系是

准确的。

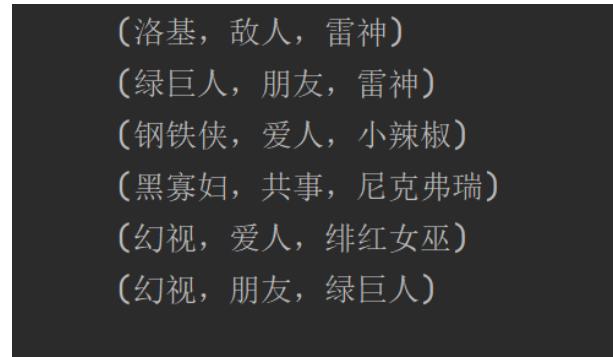


图 6 抽取范例展示

2.3 类别推断

知识图谱中，类别信息 (Type Information) 是一种特殊的三元组，其具体表现为：

$$instance \xrightarrow{type} concept$$

在我们组构建的漫威人物知识图谱中，由于这一任务属于特定领域的知识构建，所以我们根据类别推断的方法和目标进行一定的延伸，使用类别推断来进行漫威人物所属的组织的判断，具体的类别包括复仇者联盟、神盾局特工、银河护卫队、超级反派等。

在知识图谱构建中，我们结合了两种类别信息推断的方式，包括

- Type Inference from Infoboxes
- Type Inference from Text

2.3.1 Type Inference from Infoboxes

分析从百度百科得到的 infobox，部分漫威人物的 infobox 中有“所属团队”这一属性，如图7所示。对于有该属性的人物，我们通过爬取的网页信息，用 xpath 解析后使用正则表达式匹配这一属性，抽取出其属性值，即该人物所属团队。

图 7 Type Inference from Infoboxes

2.3.2 Type Inference from Text

部分人物的 infobox 中不存在“所属团队”这一属性。对于这些人物，我们选择从文本中进行类别提取，即通过识别轻量级语法模式进行提取。

从百度百科提取的人物 infobox 中，我们可以从人物经历中推断如否加入了诸如神盾局特工、复仇者联盟等团队，如 A 加入了 B（复仇者联盟、神盾局特工），其判断的正则表达式如下：

$$pattern = re.compile('(.*)[\u52a0][\u5165](.*)[\u590d][\u4ec7][\u8005](.*)$')$$

图 8 Type Inference from Text (1)

对于反派人物，其数据中的 lemma-summary 中都会包含语句“漫威漫画旗下超级反派”，如图9所示，据此进行正则匹配推断是否属于反派。

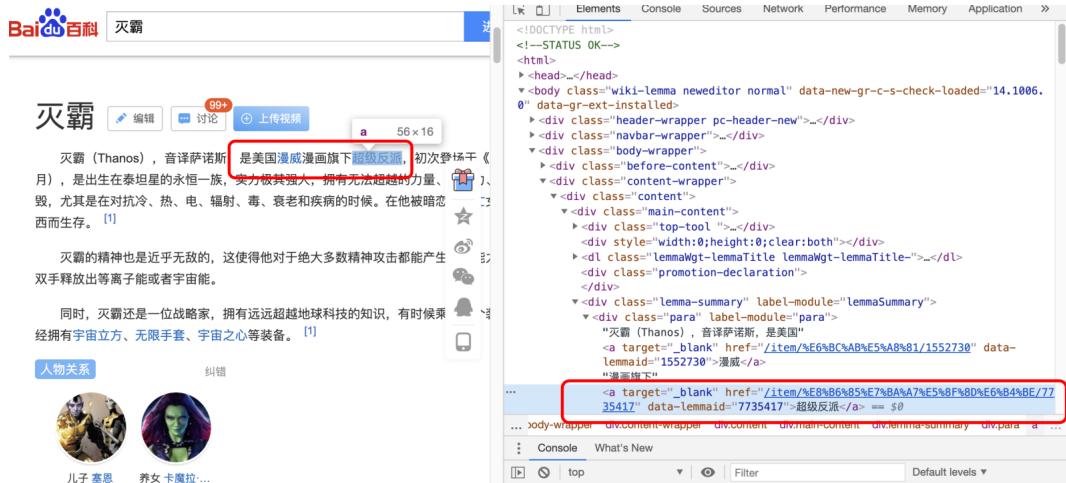


图 9 Type Inference from Text (2)

根据上述的方法，首先我们对需要进行类别推断的漫威人物爬取其百科信息，使用 python 包 xpath 解析网页文件，判断其基本信息中是否有属性“所属团队”，有则将其属性值作为团队类别；若没有这一属性，则匹配 lemma-summary 中是否有“漫威旗下超级反派”，若有则所属类别为反派，没有则对其 infobox 的属性“人物经历”的文本进行语法模式提取，提取出加入的团队类别。

对所有数据使用上述算法处理后，对没有不符合上述类别的归类为其他，并对得到的类别信息进行数据清洗和检查，去除了百科文本中的几个错别字并进行同义词合并，得到五个类别，对类别进行编码后保存到 json 文件中，后续将使用该信息在 Neo4j 中进行可视化。

```

"data": [
  {
    "name": "钢铁侠",
    "category": 1
  },
  {
    "name": "美国队长",
    "category": 1
  },
  {
    "name": "绿巨人",
    "category": 1
  },
  {
    "name": "雷神",
    "category": 1
  },
  {
    "name": "黑寡妇",
    "category": 1
  },
  {
    "name": "战争机器",
    "category": 1
  },
  {
    "name": "鹰眼",
    "category": 1
  },
  {
    "name": "尼克弗瑞",
    "category": 2
  }
]

```

图 10 Results of type inference

2.4 知识融合

知识融合目标是融合各个层面（概念层、数据层）的知识，基本的问题都是研究怎样将来自多个来源的关于同一个实体或概念的描述信息融合起来

2.4.1 数据预处理

数据预处理阶段，原始数据的质量会直接影响到最终链接的结果，不同的知识库对同一实体的描述方式往往是不相同的，对这些数据进行归一化是提高后续链接精确度的重要步骤，我们用 beautifulsoup 解析爬取的网页，过滤掉多余信息，只保留 infobox，基本信息等数据，整合为 key-value 的形式并保存在 json 文件中。

得到 json 文件后，英文，注释，网页中原有的错误，空格，符号等问题使用正则表达式和手动替换的方式调整。

图 2 中左侧为从维基百科中爬取的数据，右侧为从百度百科中爬取的数据

wiki_raw.json > {} 美国队长	data.json > {} 贝蒂布兰特
1 {	1 {
2 "美国队长": [2 "钢铁侠": [
3 {"基本信息": "美国队长",	3 "中文名": "安东尼·爱德华·"托尼"·斯塔克",
4 "出版信息": "出版商漫威漫画首次登场《Captain America Comics》",	4 "外文名": "Anthony Edward "Tony" Stark",
5 "创作者": "乔·西蒙, 杰克·科比",	5 "别名": "托尼·史塔克、Iron Man (钢铁侠)",
6 "故事信息": "",	6 "饰演": "小少爷特·唐尼",
7 "真名": "史蒂芬·罗杰斯",	7 "配音": "小少爷特·唐尼 (原声带)",
8 "种族": "人类",	8 "性别": "男",
9 "所属团队": "复仇者联盟神盾局美国陆军秘密复仇者, 侵略者联盟, Captain's Unnnamed",	9 "登场作品": "电影: 《钢铁侠》《钢铁侠》《无敌浩克》(客串)《钢铁侠2》《复仇者联盟》",
10 "伙伴关系": "巴基, 娜塔丽·卡特, 漆魔, 黑寡妇, 红女巫",	10 "生曰": "1970年5月29日",
11 "能力": "强化的体能徒步与持械的格斗专家高明的战术家和擅场司令官使用合金盾牌",	11 "虚拟人物血型": "O型",
12 }],	12 "身高": "本人: 6英尺1英寸 (185cm); 穿上战衣: 6英尺6英寸 (198cm)",
13 "绿巨人": [13 "体重": "本人: 225磅 (102kg); 穿上战衣: 425磅 (193kg)",
14 {"基本信息": "红女巫红巫女, 艺术家为Frank Cho绘制",	14 "逝世时间": "2023年10月中旬",
15 "出版信息": "出版商漫威漫画首次登场《X战警》",	15 "毕业院校": "麻省理工学院",
16 "创作者": "斯坦·李杰克·科比",	16 "所属企业": "斯塔克工业",
17 "故事信息": "",	17 "主要成就": "钢铁侠"Mark"系列",
18 "真名": "旺达·马克西莫夫",	18 "所属团队": "复仇者联盟、神盾局、光棍会",
19 "种族": "人类",	19 "经历": "钢铁侠 (Iron Man) 是美国漫威漫画旗下的超级英雄, 有多代钢铁侠, 且",
20 "重要别名": "非凡复仇者, 变种人兄弟情, 复仇者联盟女性解放者联盟, 西海岸复仇者",	20 "能力": "官方数据智力6力量2 (着装装甲时为6) 速度2 (着装装甲时为5) 耐力1",
21 "能力": "幻象魔法泡泡魔法扭曲现实控制几率",	21 "国力": "美国队长 (Captain America) 是美国漫威漫画旗下招摇英雄, 初次登上",
22 },	22 "美国队长": [
23 "雷神": [23 "中文名": "史蒂夫·罗杰斯",
24 {"基本信息": "雷神",	24 "外文名": "Steve Rogers",
25 "出版信息": "出版商漫威漫画首次登场《神秘之旅》",	25 "别名": "Captain America (美国队长、美国上尉)",
26 "创作者": "斯坦·李拉里·伯杰·科比基于神话人物",	26 "国籍": "美国",
27 "故事信息": "全名托尔·奥丁森",	27 "民族": "爱尔兰人, [1]",
28 "种族": "阿萨神族原住地阿斯加德",	28 "出生地": "美国-纽约-布鲁克林区",
29 "所属团队": "阿斯加德, 复仇者联盟三勇士, 托尔军团",	29 "出生日期": "1918年7月4日, [2]",
30 "重要别名": "Sigsteinn, Siegrie, Dr. Donald Blak, Jake Oslo, Sigurd Jarl",	30 "身高": "6英尺2英寸 (188cm)",
31 "能力": "超人的力量, 速度, 力耐和感官, 不需空气, 食物, 水等生存基本条件心灵攻击",	31 "体重": "240磅 (108kg)",
32 },	32 "职业": "军人",
33 "黑寡妇": [33 "代表作品": "《美国队长》系列、《复仇者联盟》系列",
34 {"基本信息": "黑寡妇 (黑寡妇) 第1期封面由丹尼尔·阿库尼尼绘制",	34 "主要成就": "领导复仇者联盟和复仇者联盟、美利坚诸位超级英雄的精神领袖",
35 "出版信息": "出版商漫威漫画首次登场《悬疑故事》第52期",	35 "所属团队": "漫威、全能战队、复仇者联盟、神盾局",
36 "创作者": "斯坦·李唐·里科唐·赫克",	36 "主要能力": "远超常人级别的各项体能",
37 },	37 "武器装备": "振金与埃德曼合金制成的盾牌",
38 "创作者": "斯坦·李唐·里科唐·赫克",	38 "经历": "美国队长 (Captain America) 是美国漫威漫画旗下招摇英雄, 初次登上",
39 "故事信息": "",	39 "黑寡妇": [
40 "真名": "娜塔利娅·艾丽安诺芙娜·罗曼诺娃Natalia Alianova Romanova",	40 "中文名": "罗伯特·布雷思·班纳",
41 "种族": "人类",	41 "外文名": "Robert Bruce Banner",
42 "所属团队": "神盾局, 复仇者联盟, 冠军队, 克格勃, 强大复仇者, 女性解放者联盟, 复仇者联盟",	42 "别名": "超级英雄",
43 "伙伴关系": "美国队长, 夜魔侠, 鹰眼, 冬兵",	43 "演绎": "爱德华·诺顿、马克·鲁法洛",
44 "重要别名": "娜塔莉·拉什曼, 佛拉·瑞纳斯, 琳丽·法瑞尔, 娜塔莎·罗曼诺夫, 十月",	44 "登场作品": "《绿巨人》系列、《复仇者联盟》系列",
45 "能力": "军事战术专家, 使徒搏斗高手, 秘密特工; 衰老速度减缓, 免疫系统加强, 战争机器",	45 "身高": "变身后: 5英尺6英寸 (175cm) / 变身前: 7英尺6英寸 (229cm)",
46 },	46 "体重": "变身后: 128磅 (59kg) / 变身前: 115磅 (52kg)",
47 "战争机器": [47 "出生地": "美国-俄亥俄州-代顿",
48 {"基本信息": "战争机器",	48 "国籍": "美国",
49 "出版信息": "出版商漫威漫画首次登场以詹姆斯·罗德斯身份: 《钢铁侠》8以钢铁侠",	49 "经历": "罗伯特·布鲁斯·班纳 (Robert Bruce Banner) 即绿巨人浩克 (Hulk)",
50 "创作者": "詹姆斯·罗德斯",	50 "能力": "智力2 (班纳人格为6) 力量7速度3耐力7能量发射1 (伽马能量失控时为5)",
51 "真名": "詹姆斯·'罗德斯'"罗德斯",	51 "雷神": [
52 "种族": "人类",	52 "中文名": "雷神·托尔·奥丁森",
53 "所属团队": "美国陆军陆战队, 斯塔克工业, 复仇者, 钢铁军团, 秘密复仇者, 美国国际",	53 "外文名": "Thor / Thor Odinson",
54 "重要别名": "钢铁侠钢铁爱国主义者",	54 "别名": "索尔·托尔",
55 "能力": "能熟练操作各种常规武器熟练的军用格斗技巧专业的航空学知识熟练的飞行技能",	55 "国籍": "阿萨神族",
56 },	56 "民族": "阿萨神族",
57 "鹰眼": [57 "主要成就": "参与创立复仇者联盟、结束诸神黄昏混沌、击败奥丁兄长大蛇库尔、摧毁山洞",
58 {"基本信息": "鹰眼2004年4月的鹰眼第5期封面, 由卡洛斯·帕切科及杰西·美利所绘",	58 "出生地": "瑞典",
59 "出版信息": "出版商漫威漫画",	59 "代表作品": "《雷神》、《雷神2》、《复仇者联盟》系列",
60 "创作者": "斯坦·李, don Heck",	60 "涉及领域": "神话、漫画、动画、电影",
61 "故事信息": "",	61 "所属组织": "复仇者联盟",
62 "真名": "米尔斯·弗朗西斯·'巴顿",	62 "主要能力": "超级力量、超级速度、超级耐力、操控雷电和风力、使用仙宫魔法",
63 "种族": "人类",	63 "附加属性": "超级力量和奥丁之力的继承者, 雷神之锤姆吉尔尼尔的掌控者",
64 "所属团队": "复仇者联盟复仇者学院, 秘密复仇者, 挡箭者, 新复仇者",	64 "身高": "198公分 (6英尺6英寸)",
65 "重要别名": "黑寡妇防身鸟斯特·毕易普",	65 "体重": "290千克 (640磅)",
66 "美国队长": "美国队长, 歌利亚, 猛人",	66 "父母亲": "生神奥丁和大地女神盖亚",
67 "尼克弗瑞": [67 "养父母": "弗丽嘉",
68 {"基本信息": "尼克·弗瑞",	68 "兄弟姐妹": "提尔·巴德尔、海达尔、洛基",
69 "出版信息": "出版商漫威漫画首次登场《Sgt. Fury and his Howling Commandos》",	69 "好友": "九界中善良的生物、复仇者们、北欧神话信徒。其它神话中的英雄",
70 "创作者": "杰克·科比, 斯坦·李",	70 "经历": "托尔·奥丁森 (Thor Odinson), 即雷神, 是美国漫威漫画旗下的超级英雄",
71 "故事信息": "全名尼古拉斯·约瑟夫·"尼克"·弗瑞",	71 "能力": "智力3力量7速度7耐力7能量发射e战斗技能4",
72 "所属团队": "Secret Warriors, M.I.T.A., 美国陆军C.I.A., 哆略突击队, Great White",	72 "中文名": "娜塔莎·罗曼诺夫",
73 "重要别名": "Scorpi, Gemini, various others on undercover missions",	73 "外文名": "Natalia Alianova Romanova, Natasha Romanoff",
74 }	74 "国籍": "俄罗斯",
75 },	75 "种族": "人类",
76 }	76 "民族": "俄罗斯",
77 }	77 "经历": "娜塔莎·罗曼诺夫 (Natasha Romanoff) 是美国漫威漫画旗下招摇英雄, 初次登上",

图 11 维基百科, 百度百科, infobox 数据对比

由于中文维基百科的数据较少, 所以我们的知识图谱以百度百科的数据为主体构建, 将维基百科的数据融合进百度百科。

2.4.2 数据特征分析

中文名: "娜塔莎·罗曼诺夫",
"外文名": "Natalia Alianova Romanova, Natasha Romanoff",
"别名": "Black Widow (黑寡妇) 、娜塔莉亚·艾丽安诺芙娜·罗曼诺娃",
"饰演": "斯嘉丽·约翰逊、海蒂·门尼梅可",
"性别": "女",
"登场作品": "《黑寡妇》系列、《复仇者联盟》系列",
"身高": "1.7 m",
"体重": "59 公斤",
"国籍": "前苏联",

```
        "加入团队": "复仇者联盟、秘密复仇者、神盾局",
        "上一任黑寡妇": "克莱尔·瓦扬",
        "下一任黑寡妇": "叶莲娜·贝洛娃",
        "经历": "...",
        "能力": "智力3力量3速度2耐力3能量发射3战斗技能6"
    }

{

    "基本信息": "黑寡妇《黑寡妇》第1期封面由丹尼尔·阿库尼亚绘制",
    "出版信息": "出版商漫威漫画首次登场《悬疑故事》第52期",
    "创作者": "斯坦·李, 唐·里科唐·赫克",
    "故事信息": "",
    "真名": "纳塔利娅·爱丽安诺芙娜·罗曼诺娃",
    "种族": "人类",
    "所属团队": "神盾局, 复仇者联盟, 冠军队, 克格勃, 强大复仇者, 女性解放者联盟, 雷霆特攻队, 秘密复仇者, 雇佣英雄团",
    "伙伴关系": "美国队长, 夜魔侠, 鹰眼, 冬兵",
    "重要别名": "娜塔莉·拉什曼, 劳拉·玛瑞斯, 玛丽·法瑞尔, 娜塔莎·罗曼诺夫, 十月, 叶莲娜·贝洛娃",
    "能力": "军事战术专家, 徒手搏斗高手, 秘密特工; , 衰老速度减缓, 免疫系统加强; , 专家级射击手; 精通各类武器; , 多功能战
}
```

2.4.3 基于半监督的规则学习算法

将百度百科, 维基百科的实体, 转换成三元组的格式, 每行一个三元组, 分别表示实体标题/属性/属性值, 以 6 个分号间隔,

1	美国队长;;;;;中文名;;;;;史蒂夫·罗杰斯	1	美国队长;;;;;出版信息;;;;;出版商漫威漫画，首次登场《CaptainAmericaComics》
2	美国队长;;;;;外文名;;;;;SteveRogers	2	美国队长;;;;;创作者;;;;;乔·西蒙，杰克·科比
3	美国队长;;;;;别名;;;;;CaptainAmerica (美国队长, 美国上尉)	3	美国队长;;;;;种族;;;;;人类
4	美国队长;;;;;国籍;;;;;美国	4	美国队长;;;;;所属团队;;;;;复仇者联盟神盾局美国陆军秘密复仇者，便略者联盟，CaptainAmericaTeam
5	美国队长;;;;;民族;;;;;爱尔兰人, [1]	5	美国队长;;;;;伙伴关系;;;;;巴基, 佩姬·卡特, 猛鹰, 黑寡妇, 红女巫
6	美国队长;;;;;出生地;;;;;美国-纽约-布鲁克林区	6	美国队长;;;;;能力;;;;;强化的体能使得手与持械的格斗专家高明的战术和控场司令官使用
7	美国队长;;;;;出生日期;;;;;1918年7月4日, [2]	7	美国队长;;;;;出版信息;;;;;出版商漫威漫画首次登场《X战警》
8	美国队长;;;;;身高;;;;;6英尺2英寸 (198cm)	8	绿巨人;;;;;创作者;;;;;斯坦·李, 杰克·科比
9	美国队长;;;;;体重;;;;;240磅 (108kg)	9	绿巨人;;;;;种族;;;;;人类
10	美国队长;;;;;职业;;;;;军人	10	绿巨人;;;;;所属团队;;;;;非凡复仇者，变种人兄弟帮，复仇者联盟，女性解放者联盟，西
11	美国队长;;;;;代表作品;;;;;《美国队长》系列、《复仇者联盟》系列	11	绿巨人;;;;;重要别名;;;;;旺达·弗兰克, 安娜·马克斯费夫, 旺达·马格努斯
12	美国队长;;;;;主要成就;;;;;领导侵袭者和复仇者联盟、美国诸位超级英雄的精神领袖	12	绿巨人;;;;;能力;;;;;幻象魔法混沌魔法扭曲现实控制几率
13	美国队长;;;;;所属团体;;;;;侵袭者，全胜战队，复仇者联盟，神盾局	13	雷神;;;;;出版信息;;;;;出版商漫威漫画首次登场《神秘之旅》
14	美国队长;;;;;主要能力;;;;;远超常人极限的各项体能	14	雷神;;;;;创作者;;;;;斯坦·李拉里·李伯杰·科比基于神话人物
15	美国队长;;;;;武器装备;;;;;振金与埃德曼合金制成的盾牌	15	雷神;;;;;种族;;;;;阿斯神族原居地阿斯加德
16	美国队长;;;;;智力;;;;;智力4速度3耐力3能量发射1战斗技能7	16	雷神;;;;;所属团队;;;;;阿斯加德，复仇者联盟三勇士，托尔军团
17	绿巨人;;;;;中文名;;;;;罗伯特·布鲁斯·班纳	17	雷神;;;;;重要别名;;;;;Siegman, Siegfried, Dr.DonaldBlk, JakeOlso, SigurdJarl
18	绿巨人;;;;;别名;;;;;超级英雄	18	雷神;;;;;能力;;;;;超人的力量，速度，耐力和感官，不需空气，食物，水等生存基本条件
19	绿巨人;;;;;饰演;;;;;漫德华·诺顿、马克·鲁法洛	19	黑寡妇;;;;;出版信息;;;;;出版商漫威漫画首次登场《悬疑故事》第52期
20	绿巨人;;;;;登场作品;;;;;《绿巨人》系列、《复仇者联盟》系列	20	黑寡妇;;;;;创作者;;;;;斯坦·李唐·黑科特·赫克
21	绿巨人;;;;;身高;;;;;变身后: 1英尺6英寸 (175cm) / 变身后: 7英尺6英寸 (229cm)	21	黑寡妇;;;;;种族;;;;;人类
22	绿巨人;;;;;体重;;;;;变身前: 128磅 (58kg) / 变身后: 1150磅 (522kg)	22	黑寡妇;;;;;所属团队;;;;;神盾局，复仇者联盟，冠军队，克格勃，强大复仇者，女性解放
23	绿巨人;;;;;出生地;;;;;美国-俄亥俄州-代顿	23	黑寡妇;;;;;伙伴关系;;;;;美国队长，夜魔侠，鹰眼，冬兵
24	绿巨人;;;;;国籍;;;;;美国	24	黑寡妇;;;;;重要别名;;;;;娜塔莉·拉什曼, 劳拉·玛瑞斯, 玛丽·法瑞尔, 娜塔莎·罗曼诺夫
25	绿巨人;;;;;能力;;;;;智力2力量2耐力3能量发射1	25	黑寡妇;;;;;能力;;;;;军事战术专家，徒手搏斗高手，秘密特工，衰老速度减缓，免疫系统
26	雷神;;;;;中文名;;;;;托尔·奥丁森		
27	雷神;;;;;外文名;;;;;Thor / ThorOdinson		
28	雷神;;;;;别名;;;;;索尔、托尔		
29	雷神;;;;;国籍;;;;;阿斯加德		
30	雷神;;;;;民族;;;;;阿斯神族		
31	雷神;;;;;主要成就;;;;;参与创立复仇者联盟、结束诸神黄昏循环、击败奥丁兄长大蛇库尔、有奥丁之力的情况下单独击败灭霸和漫威戈		
32	雷神;;;;;代表作品;;;;;《雷神》, 《雷神2》, 《复仇者联盟》系列		
33	雷神;;;;;涉及领域;;;;;神话、漫画、动画、电影		
34	雷神;;;;;所属组织;;;;;复仇者联盟		

图 12 百度百科，维基百科，实体标题/属性/属性值三元组

我们定义 Levenshtein Distance 编辑距离计算属性相似度，即最小编辑距离，目的是用最少的编辑操作将一个字符串转换成另一个，Levenshtein Distance 的定义如下，其中，+1 表示的是插入，删除和替换操作的代价。

$$D(i, j) = \begin{cases} D(0, 0), & 0 \\ D(i, 0), & D(i - 1, 0) + 1 \\ 1 < i < N \\ D(0, j), & D(0, j - 1) + 1 \\ 1 < j < M \\ & \begin{cases} D(i - 1, j) + 1 \\ D(i, j - 1) + 1 \\ D(i - 1, j - 1) + 1 \end{cases} \end{cases}$$

我们将维基百科和百度百科的实体数据输入到模型中，生成等价实体

钢铁侠	钢铁侠
美国队长	美国队长
绿巨人	绿巨人
雷神	雷神
黑寡妇	黑寡妇
战争机器	战争机器
鹰眼	鹰眼
尼克弗瑞	尼克弗瑞
猎鹰	猎鹰
幻视	幻视
绯红女巫	绯红女巫
快银	快银
奥创	奥创
特工希尔	特工希尔
特工卡特	特工卡特
海姆达尔	海姆达尔
小辣椒	小辣椒
霍华德·史塔克	霍华德·史塔克
玛丽亚·斯塔克	玛丽亚·斯塔克
埃德温·贾维斯	埃德温·贾维斯
奥丁	奥丁
希芙	希芙
特查拉	特查拉
灭霸	灭霸
星爵	星爵
卡魔拉	卡魔拉
毁灭者德拉克斯	毁灭者德拉克斯
火箭浣熊	火箭浣熊
格鲁特	格鲁特
勇度	勇度
星云	星云
罗南	罗南
黄蜂女	黄蜂女
伊戈	伊戈
银影侠	银影侠
蚁人	蚁人
初代蚁人	初代蚁人
黄蜂侠	黄蜂侠
奇异博士	奇异博士
莫度男爵	莫度男爵
古一	古一
彼得·帕克	彼得·帕克
梅·帕克	梅·帕克
闪电·汤普森	闪电·汤普森
乌木喉	乌木喉
黑矮星	黑矮星
漫威之父斯坦李	漫威之父斯坦李

图 13 等价实体

接下来，我们定义了每个数据源内部的同义词

```
{
    中文名;;;;;真名
    别名;;;;;重要别名
    经历;;;;;故事信息
    加入团队;;;;;所属团队
    能力;;;;;能力
}
```

将等价的实体属性和属性值进行合并，得到知识融合后的 infobox



中文名	史蒂夫·罗杰斯
外文名	Steve Rogers
别名	Captain America (美国队长、美国上尉)
国籍	美国
民族	爱尔兰人、[1]
出生地	美国-纽约-布鲁克林区
出生日期	1918年7月4日、[2]
身高	6英尺2英寸 (188cm)
体重	240磅 (108kg)
职业	军人
代表作品	《美国队长》系列、《复仇者联盟》系列
主要成就	领导侵袭者和复仇者联盟、美国诸位超级英雄的精神领袖
所属团队	复仇者联盟，神盾局，美国陆军，秘密复仇者， 侵略者联盟
主要能力	强化的体能徒手与持械的格斗专家高明的战术家 和控场司令官使用钒合金盾牌具有举起雷神之锤的资格4倍的新陈代谢
武器装备	振金与埃德曼合金制成的盾牌
能力	智力4力量4速度3耐力3能量发射1战斗技能7
出版信息	出版商漫威漫画首次登场《Captain America Comics》
创作者	乔·西蒙，杰克·科比
伙伴关系	巴基，佩姬·卡特，猎鹰，黑寡妇，红女巫
种族	人类

图 14 融合后的 infobox

2.5 问答系统

智能问答系统以一问一答形式，从数据库中定位用户所需要的提问知识，通过与用户进行交互，为用户提供个性化的信息服务。我们在知识抽取任务得到的三元组关系的基础上，附加知识融合得到的 infobox 以及从百科上爬取的漫威角色图片，实现了一个问答系统。

2.5.1 分词模型

对于 KB-QA 问题，首要任务是如何将用户查询的语句转换成数据库可识别的问题，转换的同时还需要保证转换的高准确率。而高准确率的基础，则是表现良好的分词模型。

例如，当用户提出一个问题：“绿巨人的爱人是谁？”时，我们需要对文本数据进行处理。通过进行分词、词性标注等操作，我们能提取出关键字。在上面的问句中，我们通过词性标注可以得到：

[‘绿巨人/nr’，‘的/deg’，‘爱人/n’，‘是/vc’，‘谁/pn’]

类似地，用户所有的查询都可以进行同样操作。在阅读大量文献后，我们选择了两种模型进行实现。一种是被广泛使用的中文分词模型 LTP，另一种是 ACL2020 新提出的命名实体识别模型 FLAT。

LTP

LTP 是基于词性标注、依存语法分析的一种分词模型。其本质可以视作一个感知机。这种通过结构化感知器训练的经典模型在多个中文分词数据集上都取得了较好的效果。

FLAT

在我们构建的知识图谱中，有许多实体词条在日常生活中易出现歧义。如对于“美国队长”一词，我们很容易将“美国”和“队长”分成两个实体。所以，在用户查询语句中，我们应该重点关注命名实体识别问题。

FLAT 是一种表现 SOTA 的命名实体识别模型。它是一种基于汉字格结构和 Transformer 的模型。

汉字格结构思想的提出已有较长时间。汉字格结构是指我们可以将一个句子与一个词典进行匹配，得到其中的潜词，从而得到一个有向无环图，其中每个节点都是一个字符或一个潜在的字。格包括句子中的一系列字符和可能的单词。它们不是按顺序排列的，单词的第一个字符和最后一个字符决定了它的位置。

在汉字格结构的基础上，FLAT 提出了一种新的编码方式，将一句话的格之间的位置编码并扩展成平面。句子的编码再送入 transformer 进行训练。

我们在 MSRA 上进行了模型训练，模型测试准确率达到了 0.94。最终我们结合 FLAT、字典匹配、LTP 进行人物名称识别和分词，共同完成了任务。

2.5.2 语义解析

进行分词之后，我们采用简单的语义解析方法进行知识问答。这样的方法可以满足用户简单的查询。

查询语句中有蕴含有效信息的短语或词语，这些短语被称作表征短语。它们能够对应到知识库中的特定元素：实体，概念，关系。例如对于查询语句“钢铁侠的敌人是谁？”

”，我们可以将其映射为“？谁”，“钢铁侠”，“敌人”。根据语义解析的结果我们从数据库中返回所要查询的所有三元组。

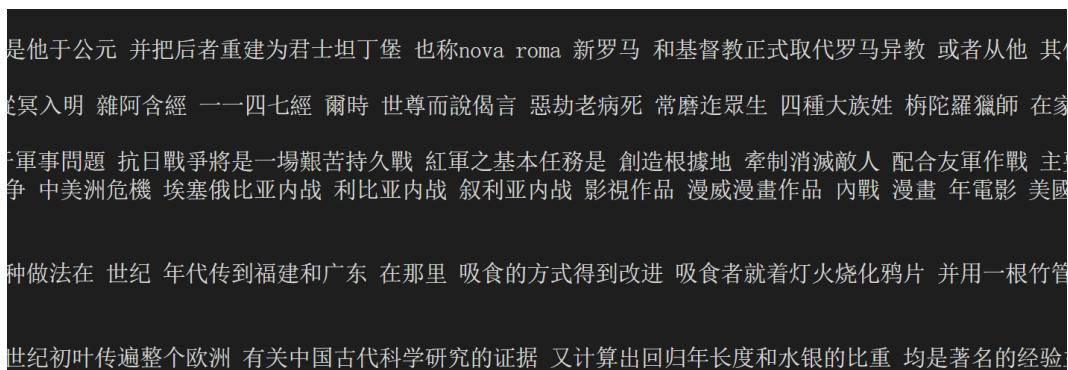
2.6 推荐系统

我们使用中文维基百科语料以及爬取到的漫威人物文本资料做了 embedding，并根据 word2vec 构建了漫威人物推荐系统，输出与用户输入人物距离最近的人物。完整处理过程包括解析中文维基百科文件，繁体简体转化，数据清洗，中文分词，Gensim 词嵌入模型训练。

对于中文分词模块，考虑人物名字以及特殊能力等名词分词存在较大难度，我们复现了 SOTA 中文分词，并与 jieba 分词的效果做了对比实验。

2.6.1 数据解析

中文维基百科数据来源于 zhwiki-20210301-pages-articles.xml.bz2 文件。先用 Gensim 库中的 WikiCorpus() 函数对其进行处理，得到繁体中文维基语料的 txt 文件。



是于公元 并把后者重建为君士坦丁堡 也称nova roma 新罗马 和基督教正式取代罗马异教 或者从他 其
至冥入明 雜阿含經 一一四七經 爾時 世尊而說偈言 惡劫老病死 常磨涅槃生 四種大族姓 梅陀羅獵師 在家
軍事問題 抗日戰爭將是一場艱苦持久戰 紅軍之基本任務是 創造根據地 牽制消滅敵人 配合友軍作戰 主
爭 中美洲危機 埃塞俄比亞內戰 利比亞內戰 叙利亞內戰 影視作品 漫威漫畫作品 內戰 漫畫 年電影 美國
种做法在 世纪 年代传到福建和广东 在那里 吸食的方式得到改进 吸食者就着灯火化鸦片 并用一根竹管
世纪初叶传遍整个欧洲 有关中国古代科学的研究的证据 又计算出回归年长度和水银的比重 均是著名的经验

图 15 繁体中文维基

对于繁体中文，我们需要将其转化为简体中文。应用 python 中的 opencc 库（Open Chinese convert）对繁体维基百科文件按行进行 t2s 处理，同时将 txt 文件中的 \t, \n 等字符进行清洗，得到中文简体维基语料。另外在分词处理前，用 iconv() 函数处理文件中的非 utf-8 字符。

2.6.2 SOTA 分词模型

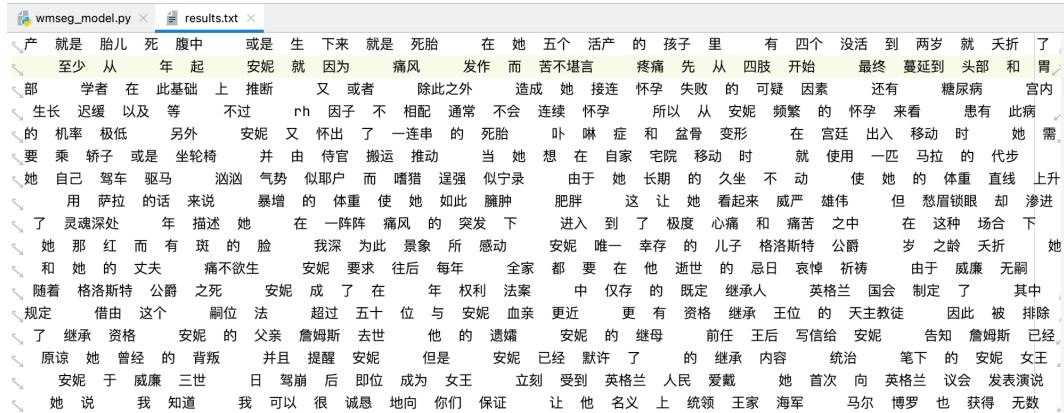
分词处理时，我们选择了 Tian 等人发表在 ACL2020 中的中文分词 SOTA 模型 WMS_{EG} ，其在 MSR、PKU 等数据集上的 OOV Recall 比之前的模型提高了约 3%。这一套分词模型在传统 encoder-decoder 模型中增加了 Memory Network。我们复现模型时，选择 ZEN 作为 encoder、crf 作为 decoder，并用项目提供的 $WMSeg.ZEN.PKU$ 预训练模

型进行分词，在应用前我们在简单语句上进行测试，分词结果良好。使用 CPU 处理完(4 核 16G macbookpro) 完整语料耗时约 2h, 分词结果如图17

```
在漫威电影宇宙中，改编为是快银的双胞胎妹妹（比快银晚出生12分钟），  
因与哥哥参与九头蛇的秘密试验并被洛基权杖上的宝石放大超能力，在打败奥创后加入复仇者联盟  
  
在漫威电影宇宙中，改编为是快银的双胞胎妹妹（比快银晚出生12分钟），  
因与哥哥参与九头蛇的秘密试验并被洛基权杖上的宝石放大超能力，在打败奥创后加入复仇者联盟
```

图 16 分词模型测试

为进行对比实验，我们同样使用了 jieba 分词对语料进行分词，jieba 分词处理耗时约 0.5h，使用两套不同分词系统的结果训练词向量，具体结果在下文展示。



```
产就是胎儿死腹中或是生下来就是死胎在她五个活产的孩子里有四个没活到两岁就夭折了  
至少从年起安妮就因为痛风发作而苦不堪言疼痛先从四肢开始最终蔓延到头部和胃  
部学者在此基础上推断又或者除此之外造成她接连怀孕失败的可疑因素还有糖尿病宫内  
生长迟缓以及等不过rh因子不相配通常不会连续怀孕所以从安妮频繁的怀孕来看患有此病  
的机率极低另外安妮又怀出了一连串的死胎卟啉症和盆骨变形在宫廷出入移动时她需  
要乘轿子或是坐轮椅并由侍官搬运推动当她想在自家宅院移动时就使用一匹马拉的代步  
她自己驾车驱马汹汹气势似耶户而嗜猎逞强似宁录由于她长期的久坐不动使她的体重直线上升  
用萨拉的话来说暴增的体重使她如此臃肿肥胖这让她看起来威严雄伟但愁眉锁眼却渗进  
了灵魂深处年描述她在一阵阵痛风的突发下进入到极度心痛和痛苦之中在这种场合下  
她那红而有斑的脸我深为此景象所感动安妮唯一幸存的儿子格洛斯特公爵岁之龄夭折她  
和她的丈夫痛不欲生安妮要求往后每年全家都要在他逝世的忌日哀悼祈祷由于威廉无嗣  
随着格洛斯特公爵之死安妮成了在年权利法案中仅存的既定继承人英格兰国会制定了其中  
规定借由这个嗣位法超过五十位与安妮血亲更近更有资格继承王位的天主教徒因此被排除  
了继承资格安妮的父亲詹姆斯去世他的遗孀安妮的继母前任王后写信给安妮告知詹姆斯已经  
原谅她曾经的背叛并且提醒安妮但是安妮已经默许了的继承内容统治笔下的安妮女王  
安妮于威廉三世日驾崩后即位成为女王立刻受到英格兰人民爱戴她首次向英格兰议会发表演说  
她说我知道我可以很诚恳地向你们保证让他名义上统领王家海军马尔博罗也获得无数
```

图 17 分词结果

2.6.3 word2vec

根据得到的分词结果，我们使用开源的第三方 Python 工具包 Gensim 进行 word2vec 模型训练，设定词向量维度为 400，模型为 CBOW 连续词袋模型。在服务器上，workers 数量为 32，训练时长约 1h。

```
model = Word2Vec(LineSentence(inp), size=400, window=5,  
min_count=5, workers=multiprocessing.cpu_count())
```

```

2021-04-08 05:53:49,045: INFO: PROGRESS: at sentence #50000, processed 46692554 words, keeping 1378183 word types
2021-04-08 05:53:51,981: INFO: PROGRESS: at sentence #60000, processed 53672165 words, keeping 1514604 word types
2021-04-08 05:53:57,922: INFO: PROGRESS: at sentence #70000, processed 60472326 words, keeping 1629968 word types
2021-04-08 05:53:57,922: INFO: PROGRESS: at sentence #80000, processed 66961122 words, keeping 1753262 word types
2021-04-08 05:54:00,577: INFO: PROGRESS: at sentence #90000, processed 73092922 words, keeping 1855850 word types
2021-04-08 05:54:03,031: INFO: PROGRESS: at sentence #100000, processed 79226974 words, keeping 1954858 word types
2021-04-08 05:54:05,483: INFO: PROGRESS: at sentence #110000, processed 85436906 words, keeping 2053646 word types
2021-04-08 05:54:07,974: INFO: PROGRESS: at sentence #120000, processed 91267850 words, keeping 2142317 word types
2021-04-08 05:54:10,283: INFO: PROGRESS: at sentence #130000, processed 96842513 words, keeping 2227957 word types
2021-04-08 05:54:12,775: INFO: PROGRESS: at sentence #140000, processed 102942987 words, keeping 2316730 word types
2021-04-08 05:54:15,010: INFO: PROGRESS: at sentence #150000, processed 108526966 words, keeping 2397976 word types
2021-04-08 05:54:17,337: INFO: PROGRESS: at sentence #160000, processed 114238947 words, keeping 2476579 word types
2021-04-08 05:54:19,625: INFO: PROGRESS: at sentence #170000, processed 119970505 words, keeping 2563362 word types
2021-04-08 05:54:21,911: INFO: PROGRESS: at sentence #180000, processed 125752037 words, keeping 2631149 word types
2021-04-08 05:54:24,226: INFO: PROGRESS: at sentence #190000, processed 131099994 words, keeping 2704073 word types

.....
2021-04-08 06:23:09,335: INFO: EPOCH 3 - PROGRESS: at 95.50% examples, 510280 words/s, in_qsize 58, out_qsize 1
2021-04-08 06:23:10,360: INFO: EPOCH 3 - PROGRESS: at 95.77% examples, 510316 words/s, in_qsize 61, out_qsize 2
2021-04-08 06:23:11,402: INFO: EPOCH 3 - PROGRESS: at 96.06% examples, 510343 words/s, in_qsize 60, out_qsize 0
2021-04-08 06:23:12,406: INFO: EPOCH 3 - PROGRESS: at 96.34% examples, 510315 words/s, in_qsize 56, out_qsize 0
2021-04-08 06:23:13,442: INFO: EPOCH 3 - PROGRESS: at 96.62% examples, 510244 words/s, in_qsize 56, out_qsize 0
2021-04-08 06:23:14,510: INFO: EPOCH 3 - PROGRESS: at 96.94% examples, 510305 words/s, in_qsize 46, out_qsize 0
2021-04-08 06:23:15,525: INFO: EPOCH 3 - PROGRESS: at 97.22% examples, 510287 words/s, in_qsize 43, out_qsize 2
2021-04-08 06:23:16,552: INFO: EPOCH 3 - PROGRESS: at 97.55% examples, 510298 words/s, in_qsize 47, out_qsize 1
2021-04-08 06:23:17,553: INFO: EPOCH 3 - PROGRESS: at 97.89% examples, 510397 words/s, in_qsize 34, out_qsize 1
2021-04-08 06:23:18,565: INFO: EPOCH 3 - PROGRESS: at 98.22% examples, 510374 words/s, in_qsize 39, out_qsize 0
2021-04-08 06:23:19,578: INFO: EPOCH 3 - PROGRESS: at 98.50% examples, 510234 words/s, in_qsize 49, out_qsize 4
2021-04-08 06:23:20,604: INFO: EPOCH 3 - PROGRESS: at 98.82% examples, 510165 words/s, in_qsize 57, out_qsize 0
2021-04-08 06:23:21,608: INFO: EPOCH 3 - PROGRESS: at 99.14% examples, 510210 words/s, in_qsize 60, out_qsize 0
2021-04-08 06:23:22,629: INFO: EPOCH 3 - PROGRESS: at 99.48% examples, 510228 words/s, in_qsize 62, out_qsize 1
2021-04-08 06:23:23,412: INFO: worker thread finished; awaiting finish of 31 more threads
2021-04-08 06:23:23,520: INFO: worker thread finished; awaiting finish of 30 more threads

```

图 18 word2vec 训练过程

训练完成后得到 60MB 的模型文件和 4.36GB 的词向量文件。加载模型文件进行测试，输入一漫威人物，得到最接近的人物，并基于此构建漫威人物推荐系统。同时我们输入三个漫威漫画旗下的超级英雄以及一个 DC 漫威旗下的超级英雄，调用 word2vec 的模型能正确进行区分，结果如图19所示（蝙蝠侠是 DC 漫画中的人物）。

```

: import gensim
model = gensim.models.Word2Vec.load("wiki.zh.text.model")
print(model.wv.doesnt_match(u"蜘蛛侠 绿巨人 黑寡妇 快银 蝙蝠侠".split()))
蝙蝠侠

```

图 19 角色所属漫画判断

2.6.4 模型分析

对于推荐得到的结果，我们发现返回的相似度排序得到的名词中除了漫威人物外还有其他的人物，若对于我们的推荐系统用户需求为推荐最相似的漫威人物，则需要对 embedding 模型进行一定调整。现有效果最好的 BERT 模型可以满足需求，其基于 Transformer 框架，包含了自注意力机制。但是我们 BERT 模型参数过多，在我们这种个人电脑上无法完成训练任务，因此我们选择在语料库中增加漫威相关的文本信息，相当于从语料库中进行一次权重分配使得 embedding 的结果对于漫威名词更加友好。新加入语料库的文本信息则来源于我们在构建知识图谱时爬取到的所有相关资料。

此外，我们对两种分词方法(jieba、SOTA 模型 WMSeg)得到的结果分别训练的 word2vec 模型进行漫威人物 similarity 测试，得到的部分结果如图20所示，可以发现

```

string = '幻视'
result_jieba = model_2.wv.most_similar(string)
result_sota = model.wv.most_similar(string)

result = zip(result_jieba[0:8], result_sota[0:8])
print('jieba' WMSeg\n')
for s1,s2 in result:
    name = s1[0]
    name2 = s2[0]
    print('{name:<{len}}\t'.format(name=name, len=22))

```

jieba WMSeg

毒藤女	汪达
奥创	水行侠
夜翼	夜魔侠
终极天神	猛毒
斗士	复仇者
毁灭者	绿箭侠
在续集	蜘蛛人
暴狼	曼达洛人

(a) '幻视'

```

string = '鹰眼'
result_jieba = model_2.wv.most_similar(string)
result_sota = model.wv.most_similar(string)

result = zip(result_jieba[0:8], result_sota[0:8])
print('jieba' WMSeg\n')
for s1,s2 in result:
    name = s1[0]
    name2 = s2[0]
    print('{name:<{len}}\t'.format(name=name, len=22))

```

jieba WMSeg

泰坦	hawkeye
破坏者	黑寡妇
绿灯侠	毁灭者
复仇者	夜魔侠
洛基	猛毒
首次出现于	密佛格
终极战士	viper
毁灭者	死侍

(b) '鹰眼'

图 20 两种分词的 embedding 结果对比

WMSeg 模型所推荐的相似人物更能满足我们真实情况中的要求。

三、数据库可视化

知识图谱构建完成后，我们将其进行了可视化展示。可视化展示基于 Neo4j 图形数据库。在于 Neo4j 框架下，每一个实体都以一个节点的形式呈现，整体数据以结构化数据存储在网络中。在图形数据库的基础上，我们进行了实体与实体间关系、知识问答的可视化实现。同时，为了能够更好地让知识图谱服务于用户，我们以网页的形式展示图形数据库。在前端部分，我们进行了一些 CSS 修饰工作。

首先，知识图谱网站首页如图下所示：

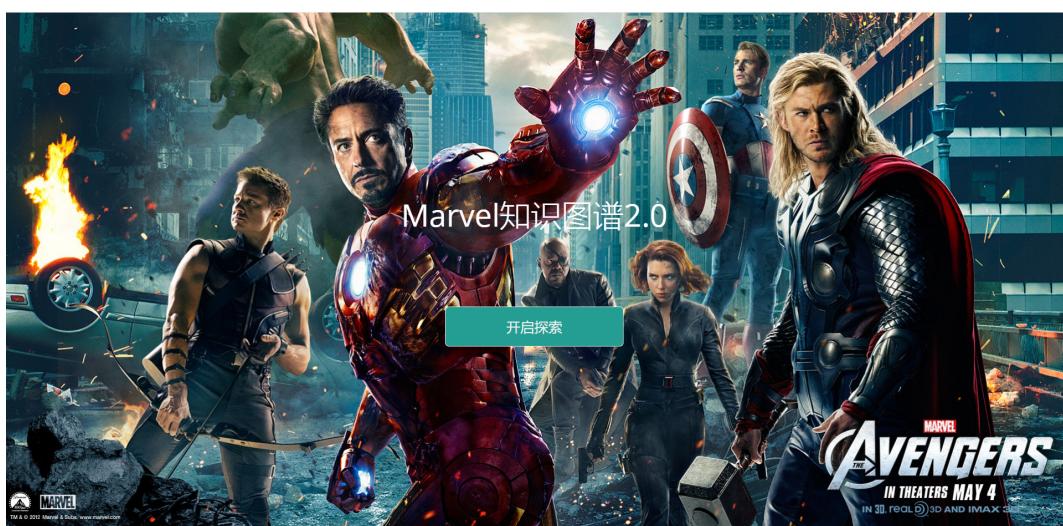


图 21 漫威知识图谱首页

进入页面之后，会有“检索人物关系”、“人物关系全貌”、“知识问答”三个模块。



图 22 模块选择

检索人物关系模块中用户可以查询到具体实体。例如用户想查询“幻视”时，在搜索栏检索“幻视”即可得到实体的与其周围实体间的联系。同时，我们在提供了热门人物的实体云展示，用户可以在实体云中选择需要查看的实体。



图 23 检索人物关系模块：实体云

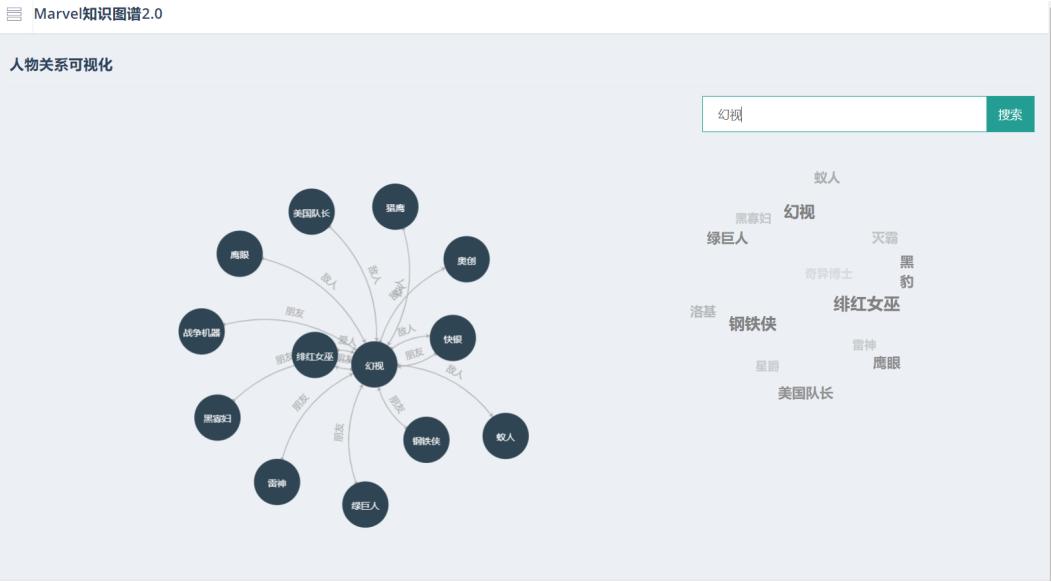


图 24 检索人物关系模块：搜索栏

人物关系全貌模块是对图形数据库中全部数据的展示，包含全部实体节点、全部实体间关系。

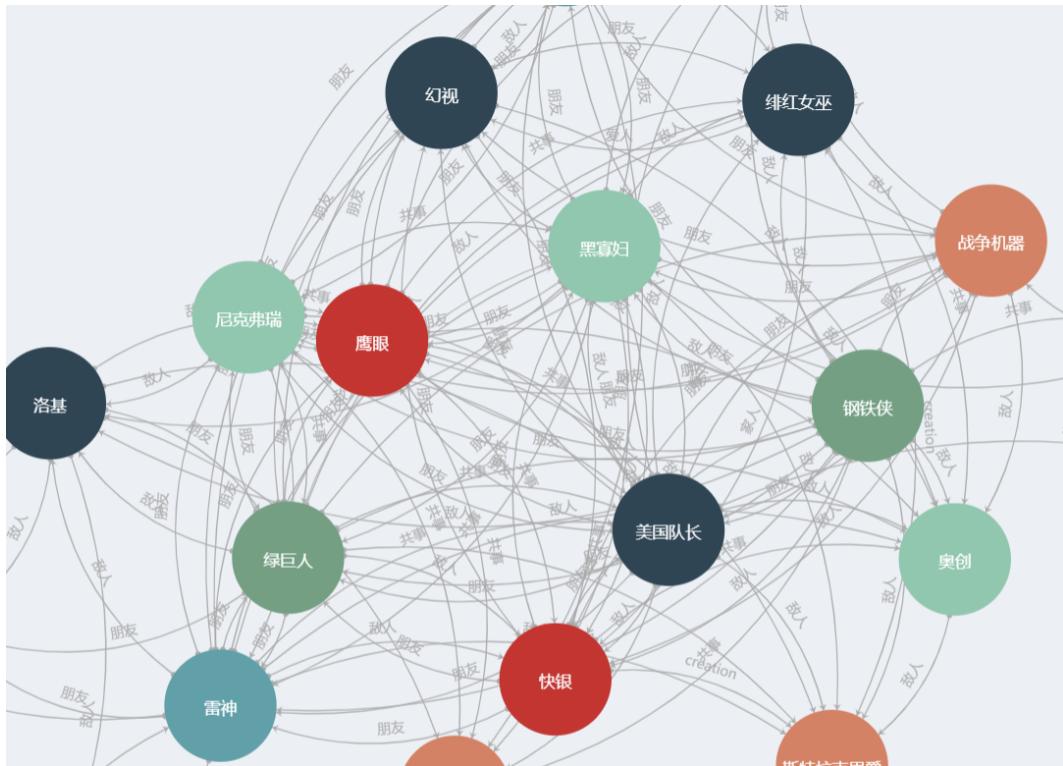


图 25 人物关系全貌

知识问答模块中，用户可以进行简单的查询。例如“黑寡妇的爱人是谁？”这一查

询，我们根据分词、语义解析的结果返回目标实体。同时，目标实体与周围节点的关系、目标实体的详细介绍都会出现在网页中，方便读者查阅。



图 26 人物关系全貌

四、心得体会

这次课程设计小组成员合力完成了特定范围人物方面的知识图谱，并据此建立了数据库和可视化界面，结合相关 NLP 算法提供查询。最大的感触是真正了解了像 Google、Baidu 这样优秀的搜索引擎背后的原理，也第一次真正意义上从头到尾自己搭建一个知识图谱的工程项目。

最开始的时候不知道知识图谱如何起步，这时候我们回想起了上学期漆桂林老师介绍过的很多工具。我们从本体构建的软件开始，设计了一个又一个本体图，考虑了人物关系中的包含与排斥。接下来又使用了 DeepKE 方法抽取人物关系。最后结合上学期所学的语义搜索内容，利用 nlp 算法做了智能问答的搜索系统。在这之后我们深刻地体会到，每一门课的内容都是醇厚的赐予，要好好利用已知的知识，也许这些知识就会在以后的任务中派上用场。

在项目中一直存在着分工问题，小组内部总共有 6 位成员，如何协调一致往前进最开始是让我们有些担心的。可随着时间的推进，小组成员们很快找到了适合自己的不同分工，哪怕是工作部分稍有冲突的情况下，也依旧能够很好地完成分工。从这个意义讲，这次实验极大地锻炼了我们的团队协作能力。

在此基础上，我们的解决问题能力也有所提高。整体框架构建的时候遇到各种各样的问题，如 neo4j 的搭建，网页爬取的不对应，有的时候甚至细微到一个 java 的版本问题。但不管问题为何，在影响团队操作的问题出现时，大家总会放下手中的工作一起解决问题。

我们遇到这样那样的问题时，有过很多次错误的尝试，但我们也在这个过程中学到了很多。在我们的小组中，其实并没有明确的组长和组员之分，每个人都很积极地在参与讨论和任务的推动，我们在分配任务的时候也不是“一言堂”，而是大家一起讨论决定，并不会让某个人特别累，而这样的讨论和任务分配的方式也让我们平时讨论的氛围更轻松，让我们推动任务的时候更认真，我们向一个真正成熟的团体越来越靠近。

总的来说，这次课程设计，除了内容知识上的学习以外，更重要的是一些团队协作能力。知识在很久以后也许会忘却，但是搜索问题、解决问题、和队友一起协商的能力，是永远不会过时的。

任务分配

组员	任务
张博宇	构建图数据库、应用 FLAT 模型
陈嘉源	类别推断、推荐系统、WMSeg 分词、外网资料爬取
王一名	事实抽取、百度百科数据爬取、DeepKE 模型训练
鲁瀚洋	知识融合、维基百科数据爬取
沈飞鸿	构建查询系统、构建分词替换字典、关系三元组抽取
廖滔	三元组关系处理

具体工作量

张博宇：调试 FLAT 模型，对于格结构长度进行调整以适应人名的提取，在 MSRA 和自建的漫威数据集上进行训练测试。基于 Neo4j 框架进行图数据库构建，把已存储的数据进行预处理后导入进 Neo4j。设计网页，提供数据节点、关系的查询，提供知识问答 api 接口。将 FLAT、LTP、替换字典整合到知识问答接口中。

陈嘉源：类别推断：从 Infobox 以及从文本中识别轻量级语法模式进行推断；推荐系统：基于中文维基百科语料和爬取的漫威人物信息训练 word2vec 模型，解析中文维基语料、繁体简体转化、中文分词、Gensim 做 word embedding；复现中文分词 SOTA 模型 WMseg 进行应用；已有漫威人物信息网页数据爬取。

王一名：本体构建，人名正确 URL 爬取，百度百科数据爬取与清洗，DeepKE 模型训练，Synonyms 中文近义词工具包实现

鲁瀚洋：维基百科数据爬取，数据清理，构建成框架所匹配的格式，对百度百科、维基百科中实体计算相似度，匹配实体，定义同义词，合并不同数据源中的属性和属性值。

沈飞鸿：构建查询系统，利用 ltp 模型完成分词，词性标注任务，进而返回查询列表（人物 + 关系），建立替换字典以解决部分分词不全的问题。使用 deepke，训练了 CNN，RNN，Transformer 三个模型从个人信息描述中抽取关系三元组。

廖滔：关系抽取数据处理。处理开源人物关系数据集，筛除掉无效的字符标点，规范化处理数字等冗余信息，并将实体对进行编码。抽取完后利用 Synonyms 中文近义词工具包将抽取到的关系转化为我们想要的“共事”“爱人”“朋友”等关系。

参考文献

- [1] Li, X. , et al. "FLAT: Chinese NER Using Flat-Lattice Transformer." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 2020.
- [2] Zhang, Y. , and J. Yang . "Chinese NER Using Lattice LSTM." (2018).
- [3] Yuanhe Tian, et al. "Improving Chinese Word Segmentation with Wordhood Memory Networks." ACL 2020
- [4] <https://zhuanlan.zhihu.com/p/39208802>
- [5] <https://radimrehurek.com/gensim/models/word2vec.html>
- [6] <https://graphics.straitstimes.com/STI/STIMEDIA/Interactives/2018/04/marvel-cinematic-universe-whos-who-interactive/index.html> (外网)
- [7] <https://github.com/zjunlp/deepke>
- [8] 《知识融合》, 漆桂林, 东南大学