

构建视频基因库

顾健 201200131030

摘要：为了提升基于内容的视频推荐的质量，有必要获取更多描述视频特征的标签，并对标签进一步加工。视频基因库也就是经过扩充和整理的影视标签库。对影视的基本信息如演员、类型等，主要是补充和纠正，对影视的情节特征，尝试通过剧情文本和关键词提炼影视在时间、地点、人物等维度上的信息，对影视的评价特征，尝试通过评论文本和关键词获取影视本身的突出特点和大众对其的普遍的感受。对应这三方面的需求，尽可能的获取足够多的分析材料，并且在处理过程中尽可能模块化、流程化。最后通过个性化栏目推荐和相关视频推荐演示视频基因库的效果。

关键字：影视标签 情节特征 评价特征 个性化栏目推荐 相关视频推荐

一、交代

视频基因库的构建本身倾向于工程，而非理论，所以即使要面对类似自然语言处理等问题，也尽量是利用已有的 NLP 工具，在提高分词、词性标注这些方面，更倾向于动用人力标注而非设计更好的算法或者模型。其他方面也类似，所以整个项目可能在理论和思想上没有创新性，在工程实现的流程和各个模块里可能会存在一些有价值的地方。

视频基因库构建之初并没有完整的思路，在具体施行过程中也因为各种原因纠正原先的计划，或者用新办法以改善效果，最后才做成现在这样，并且这和最初的构想又是有很多不同的。

视频基因库完全基于 Python3.5，数据库使用 mongoDB。使用的第三方库包括 pymongo、jieba、beautifulsoup4、tornado 等等。视频基因库是青岛海信新研发重点实验室数据智能组的项目，原始材料是数据智能组提供的影视资料，来源于华数和爱奇艺两家视频资源供应商。后期的材料主要采集自豆瓣网和 IMDB，来源于广大网民。

二、整理影视基本信息

原影视标签库 media 的基本信息存在一些问题或者瑕疵，如演员名单重复、演员排列与演员权重无关、影片类型混乱等等，为了提升影视基本标签的质量，所以决定爬取以豆瓣电影 douban 的内容为主的影视基本标签。首先要解决的问题是对应 media 的电影，如何获取相应的豆瓣电影的条目。最初的想法是将豆瓣所有的电影都爬下来，然后通过多关联匹配完成建立对应关系，确实将绝大多数的电影信息都爬到了，但是我们发现通过演员、导演、类型、时间、国家等信息进行多关联匹配的效果并不好，因为最开始 media 的内容就不是完全准确的，而且尤其是对外国人名等比较上效果没那么好，也就是可操作性没那么理想，而且最终效果大概在 50%左右的准确率，所以很快就采取了第二个办法，因为豆瓣电影有一个搜索接口，通过把 media 的 name 直接填入 url 中查询就可以获得豆瓣电影返回的可能需要的电影，简单起见直接选取了第一个作为候选的电影，人工排查后发现准确率在 80%左右，对于匹配不正确的，要不就是豆瓣电影没有相关的视频内容（media 中有很多短片、花絮等等），要不就是豆瓣检索

时将真正需要内容排在了后面。对与后者，会将策略改为爬取所有检索到的结果，再进行多关联匹配。其中，具体操作时，因为 media 对应 douban 是多对一的关系，所以爬虫的对象变得更精准了。

豆瓣电影的基本信息不仅仅质量更好，内容也更丰富，像演员和类型等等，还有演员的排列顺序等等，另外还多了更多信息，例如别名、评分等等。除了直接整理豆瓣电影的基本信息，另外还维护了知名导演表和知名演员表，在标签库的基本信息中，添加了这两个条目，这在相关影视推荐时会用得着。

三、获取影视情节特征

现在同时结合 media 和 douban 的剧情简介的内容，通过分词和词性标注后，再进行时间、地点、组织、人物、专有名词的实体识别，得到五类名词性标签，因为效果和最终实现上的考虑，最终放弃了人物和专有名词这两样内容，剩下时间、地点、组织这三类名词性标签，通过人工筛选得到了合适的内容。在爬取豆瓣信息的同时，获得了 IMDB 链接，在 IMDB 的影视介绍页面上有网民提供的关键词信息，同样获取到这些数据，并且进一步通过有道 API 完成翻译，并同样地挑出这些词中的时间、地点、组织，同样参与人工筛选。

美中不足的是，现在的名词性标签不能被足够灵活的使用，理想的场景是类似于能根据地点等信息将类似于 XXXX 年转化为比如维多利亚时代，现在还没能做到那样。通过剧情文本获取名词性标签的理由是名词能够尽可能的表达出剧情本身的一些特征或者倾向性，藏身于剧情文本的形容词所能表达的信息似乎无关痛痒。

四、获取影视评价特征

之前爬到了 douban 的短评文本的内容，通过分词和词性标注后，挑出其中的形容词，通过人工筛选得到了合适的内容。在爬取豆瓣信息的同时，获得了 IMDB 链接，在 IMDB 的影视介绍页面上有网民提供的关键词信息，同样获取到这些数据，并且进一步通过有道 API 完成翻译，并同样地挑出这些词中的名词性的词语，同样参与人工筛选。

现在形容词性标签的缺陷是，还没有建立合适的机制对短评中的形容词筛选，很多形容词具有价值但是被网民频繁的使用，同样的影片可能同时被几十个形容词描述。虽然可以接受，但是并不理想。通过评价文本获取形容词性标签的理由是形容词能够尽可能的表达出网民对影片的看法、感觉和印象，藏身于评价文本的名词所能表达的信息似乎莫名其妙。

五、基于视频基因库推荐

相对于影视标签库，视频基因库有更优质更丰富的基本信息，还有能反映剧情特征的名词性标签以及能反映观众体会和印象的形容词性标签。以这些内容为基础，应用于两种推荐场合。一种是结合用户画像的个性化栏目推荐，假设用户看过若干电影，根据电影所拥有的名词性标签和形容词性标签，找到拥有这些标签或者相近标签的其他电影，通过标签之间的组合生成个性化的栏目，以投其所好。另一种是给出一部影片，从影片类型、演员、三类名词性标签和一类形容词性标签做相关推荐。从效果出发来看，个性化栏目推荐的效果不错，但同样容易看出现有基因库的标签仍还不够方便，比如时间类的名词性标签使用不够灵活、标签组合的策略等等。相关视频推荐在一部分包含小众特征的影片中效果显著，比如解放军等等，这是最理想的情况，但这样的小众特征并不总能提取到，因此想办法进一步整理专有名词类的名词标签会对整体的效果带来很大提升，对于尚未取得小众特征的部分影片，通过主要是影视类型和演员的加权计算相似度，仍能得到比较好的效果。

六、其他

目前视频基因库的工作流程：

1. 获取华数、爱奇异的影视资料的数据文件 mediafilm
2. 根据 mediafilm 的 id、name 使用豆瓣电影的搜索接口获取对应的 doubanid、doubanname
3. 人工评估 media 影视资料和 douban 影视资料的对应质量，并优化质量
4. 考虑 media 对 douban 属于多对一，整理 douban 对 media 的一对多的列表 jointlist
5. 根据 jointlist 爬取 douban 信息，包括海报、基本信息、简介、短评、IMDB 链接等等
6. 根据 IMDB 链接获取影片关键词，使用有道翻译 api 获取翻译结果 keywords
7. 整合 media 和 douban 基本信息得到 static_b，进一步标注知名导演等得到 static_a
8. 整合 media 和 douban 影视简介通过分词、词性标注得到时间、地点、组织等名词性标签组成 segment_b，进一步过滤、映射等得到 segment_a
9. 对 douban 影视短评分词、词性标注得到形容词性的标签组成 comment_b，进一步过滤、映射等得到 comment_a
10. 对 keywords 做词性标注得到时间、地点、组织等名词性标签以及形容词性标签组成 translate_b，进一步过滤、映射等得到 translate_a
11. 对 segment_a、comment_a、translate_a 的内容按照时间、地点、组织等名词性标签以及形容词性标签分成四类，补充到 static_a 中得到新的影视资料 boot_gather，相对的，把原先 media 的影视资料整理为 boot_media
12. 对比 boot_media、boot_gather 的内容并基于 boot_gather 做个性化栏目推荐和相关影视推荐

以后视频基因库的改进方向：

1. 结合 media 数据库以改为增量式扩展
2. 爬取更多内容以增加更多的标签维度
3. 想办法获取并且描述剧情以活用标签
4. 想办法规范形容词标签不然效果不好

参考文献：

- [1] 宗成庆 《统计自然语言处理》（第2版） 清华大学出版社.
- [2] Ehud Reiter, Robert Dale 《自然语言生成系统的建造》 北京大学出版社.
- [3] 朱德熙 《语法讲义》 商务印书馆.
- [4] 项亮 《推荐系统实践》 人民邮电出版社.

效果展示：



变形金刚3

演员	希亚·拉博夫 罗西·汉丁顿-惠特莉 乔什·杜哈明 泰瑞斯·吉布森 约翰·马尔科维奇 弗兰西斯·麦克多蒙德 肯·郑 凯文·杜恩 朱丽叶·怀特 帕特里克·德姆西 艾伦·图代克		希亚·拉博夫 雨果·维文 罗西·汉丁顿-惠特莉 乔什·杜哈明 暑期观影 泰瑞斯·吉布森
名演员			
类型	动作 科幻	电影 都市 励志 动作 冒险 科幻 青春	
细类型	科幻片 动作片		
地区	美国	香港	
年份	2011	2011	
又名	变形金刚3: 黑月降临(港) 变形金刚3: 月黑之时 变形金刚 III 变3 Transformers: Dark of the Moon 3D		
打分	7.1		
标签1	迈克尔·贝 希亚·拉博夫 罗西·汉丁顿-惠特莉 乔什·杜哈明 泰瑞斯·吉布森 约翰·马尔科维奇 弗兰西斯·麦克多蒙德 肯·郑 凯文·杜恩 朱丽叶·怀特 帕特里克·德姆西 艾伦·图代克 动作 科幻 美国		英语 香港 电影 都市 励志 动作 冒险 科幻 青春 迈克尔·贝 希亚·拉博夫 雨果·维文 罗西·汉丁顿-惠特莉 乔什·杜哈明 暑期观影 泰瑞斯·吉布森 变形金刚
标签2	古代 中东 华盛顿 俄罗斯 美国 英国 意大利 墨西哥 陆战队 情报局 纯粹 宏伟 完美 有趣 壮观		当代 希望 震撼 英语 美国 救援 奋斗 香港 森林 都市 外星球 巨制 残酷 励志 暴力 世界末日 超自然 超级英雄 动作 冒险 科幻 青春 粉丝电影
时间	古代		
地点	中东 华盛顿 俄罗斯 美国 英国 意大利 墨西哥		
组织	陆战队 情报局		
评价	纯粹 宏伟 完美 有趣 壮观		

图 1 基因库与原标签库的内容比较图

☰

CLOUD


主面板

比照表

基因库

栏目推荐

相关视频



卧虎藏龙

源 > 基因库

基因库

按照标签的来源和类别展示基因库信息

☰

视频基本信息

剧情简介整合

关键词整合

评论内容整合

👆

项目	加工前	加工后
名称	卧虎藏龙	卧虎藏龙
导演	李安	李安
演员	周润发 杨紫琼 章子怡 张震 郑佩佩 郎雄 黄素影 李法曾 高西安 王德明 李黎	周润发 杨紫琼 章子怡 郑佩佩
类型	剧情 动作 爱情 武侠 古装	武侠片 动作片
地区	台湾 香港 美国 中国大陆	台湾 香港 美国 中国大陆
年份	2000	2000
别名	Crouching Tiger, Hidden Dragon	Crouching Tiger, Hidden Dragon
打分	7.7	7.7

👆 TOP

图 2 基因库基本信息加工前后的对比图

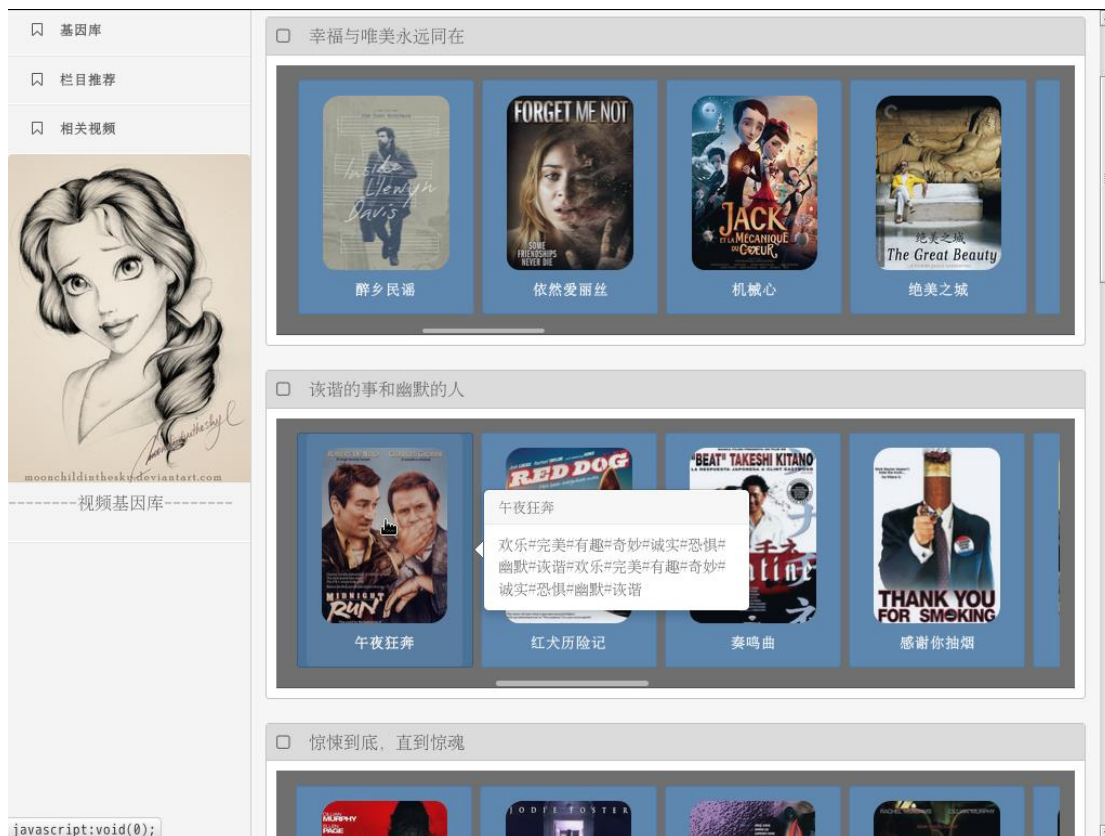


图 3 基因库应用于个性化栏目推荐效果图

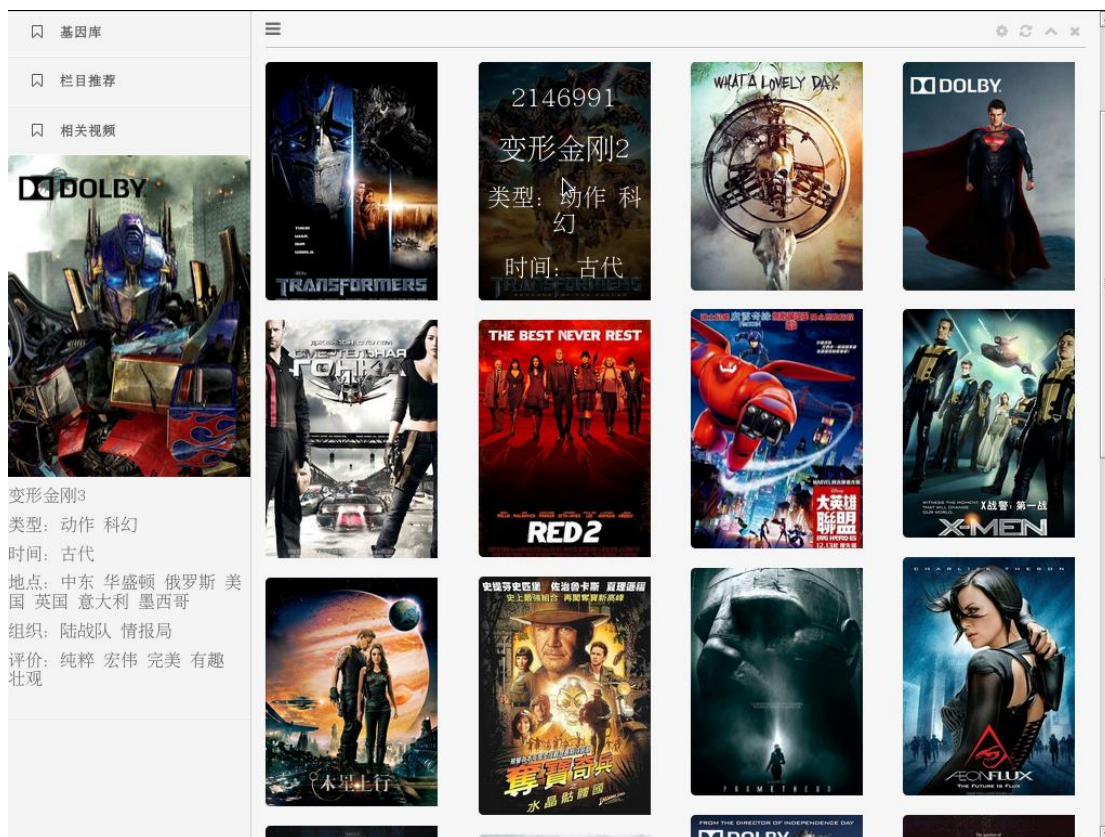


图 4 基因库应用于相关影视推荐效果图

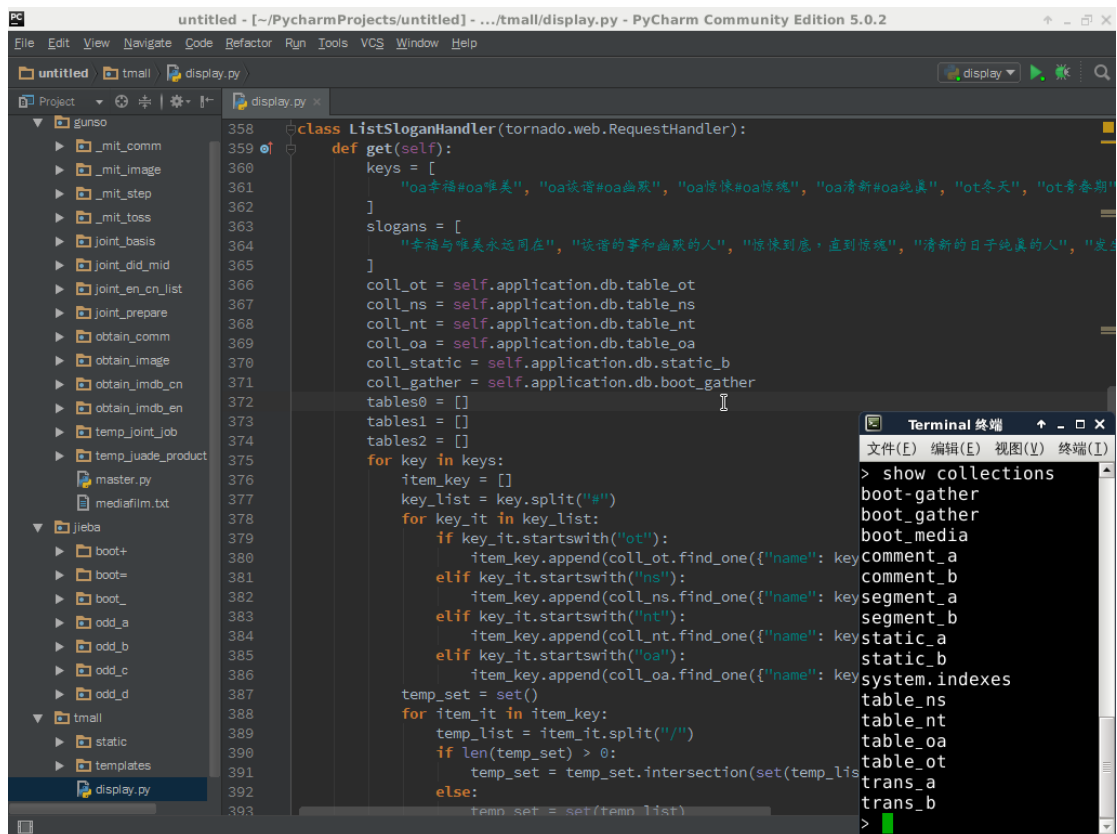


图 5 基因库项目代码和数据库文件示意图

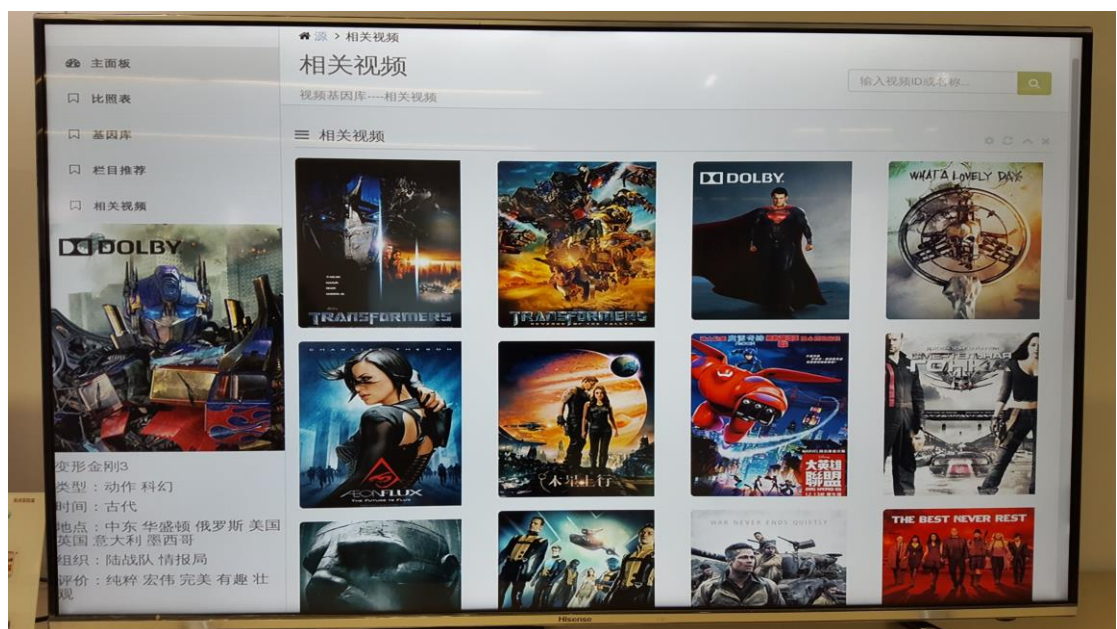


图 6 实验室开放日电视展位效果图