实验 2 豆瓣电影数据的知识感知推荐

实验背景

知识图谱（KG）在提高推荐的准确性和可解释性方面显示出了巨大的潜力。KG 中丰富的实体和关系信息可以强化用户和物品之间的关系建模，因为它们不仅揭示了物品之间的各种相关性（如两部电影由同一个人导演），还可以用来解释用户偏好（如将用户对电影的选择归因于其导演）。

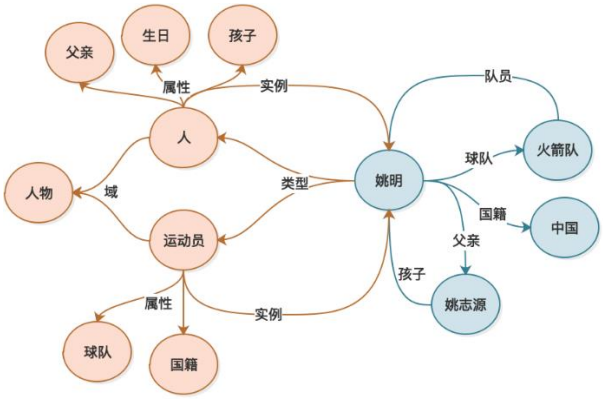
在本次实验中，我们要求各位同学从公开图谱中匹配指定电影对应的实体，并抽取合适的部分图谱，按照规则对抽取到的图谱进行处理（Stage1）；进而，基于对实验一中的豆瓣电影评分数据，结合 Stage1 所获得的图谱信息，进行可解释的、知识感知的个性化电影推荐（Stage2）。

实验介绍

Freebase 是一个由元数据组成的大型合作知识库，内容主要来自其社区成员的贡献。它整合了许多网上的资源，致力于打造一个允许全球所有人（和机器）快捷访问的资源库。Freebase 提供数据查询和录入机制。其官网（https://developers.google.com/freebase）提供 N-Triple RDF 格式的数据压缩包的下载，但请注意整个压缩包 30G，解压后 300G+。有关 Freebase 的更多信息可参考相关介绍文章[1]。

下面列举一个实例，格式为“MID、属性、值”，例如：

“<m.01jzhl>　<type.object.name>　姚明”

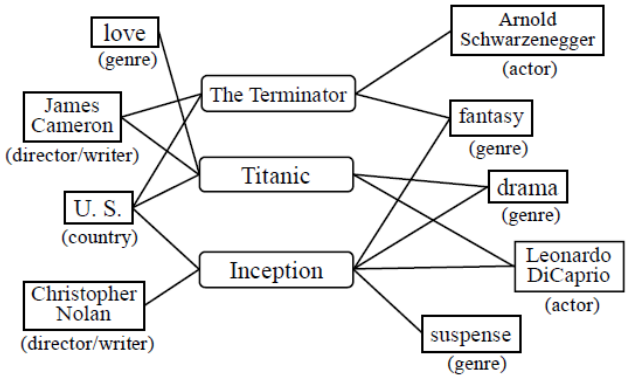


值得一提的是，Freebase 中包含了非常丰富的电影信息，这使得我们将实验一、二进行联动成为可能。在本次实验中，我们提供了实验一中给出的 1000 部豆瓣电影 ID 与 Freebase 中对应电影实体的映射关系（共涉及 578 部可匹配的电

[1] https://developer.aliyun.com/article/717320?spm=a2c6h.14164896.0.0.535f3630po4hs1

影）。实验将围绕这些实体以及其他相关信息所形成的中等规模图谱展开，分别包含图谱抽取和图谱推荐两个阶段，其中第一阶段任务详述如下：

第一阶段任务：图谱抽取



在我们给出的链接信息文件 douban2fb.txt 中，提供了豆瓣电影 ID 到图谱实体 ID 之间的映射关系，其中第一列为豆瓣电影 ID（与实验一中所提供的电影 ID 相同），第二列为 Freebase 中对应电影实体的 ID。一旦完成这样的实体链接，我们就能够借助 Freebase 抽取用于电影推荐系统的电影知识图谱。

第一阶段（Stage 1）的实验内容包含以下部分：

[1] 【必做】根据实验一中提供的电影 ID 列表，匹配获得 Freebase 中对应的实体（共 578 个可匹配实体）

[2] 【必做】以 578 个可匹配实体为起点，通过三元组关联，提取一跳可达的全部实体，以形成新的起点集合。重复若干次该步骤，并将所获得的全部实体及对应三元组合并为用于下一阶段实验的知识图谱子图。

说明及技巧：

- 三元组中应包含至少一个起点实体，无论其是作为头实体还是尾实体。
- 为保证质量，最好只保留具有< http://rdf.freebase.com/ns/前缀的实体。
- 一般而言，两跳后所形成的子图，即重复两次该步骤所获得的子图，即包含足以支撑推荐系统的丰富语义。但考虑到仅仅 578 部电影所形成的子图可能关联性较差，也可重复更多遍以获得关联性更好的子图。
- 此外，为了保障图谱的质量，也可以根据统计对不常出现的实体或关系进行筛选。例如，可以过滤掉涉及三元组少于 10 个的实体，或只保留至少在 50 个三元组中出现的关系等。

[3] 【选做】根据实验一中提供的电影 Tag 信息，在图谱中添加一类新实体（Tag 类），并建立其与电影实体的三元组，以充实电影的语义信息。

[4] 【选做】对 Tag 类实体进行实体对齐，以合并部分具有相同/高度相似语义的实体，从而精简图谱并强化其关联性。

数据集说明

我们一共提供了两个文件，包括：

中等规模图谱 freebase_movie.gz，以（头实体，关系，尾实体）这种三元组的形式进行保存，因原体积（52G）过大，采用压缩形式进行存储，如无必要请勿解压。保存形式和使用帮助如下所示，

```
<http://rdf.freebase.com/ns/m.03jt67r>  <http://rdf.freebase.com/ns/film.performance.film>  <http://rdf.freebase.com/ns/m.01vg3r>  .
<http://rdf.freebase.com/ns/m.045y4ww>  <http://rdf.freebase.com/ns/common.webpage.topic>  <http://rdf.freebase.com/ns/m.06gjk9>  .
<http://rdf.freebase.com/ns/m.0599qd8>  <http://rdf.freebase.com/ns/base.wfilmbase.siteid.film> <http://rdf.freebase.com/ns/m.033fqh>  .
<http://rdf.freebase.com/ns/m.059x0w>   <http://rdf.freebase.com/ns/film.producer.film> <http://rdf.freebase.com/ns/m.0dllxnf>  .
<http://rdf.freebase.com/ns/m.059x0w>   <http://rdf.freebase.com/ns/film.producer.film> <http://rdf.freebase.com/ns/m.047rkcm>  .
```

```python
import gzip

with gzip.open('./web/freebase_douban.gz', 'rb') as f:
    for line in f:
        line = line.strip()
        triplet  = line.decode().split('\t')
        print(triplet[0:3])
        break
```

```
['<http://rdf.freebase.com/ns/award.award_winner>', '<http://rdf.freebase.com/ns/type.type.instance>', '<http://rdf.freebase.com/ns/m.04n2x3p>']
```

链接信息文件 douban2fb.txt，提供了豆瓣电影 ID 到 Freebase 图谱实体 ID 之间的映射关系。其中第一列为豆瓣电影 ID（与实验一中所提供的电影 ID 相同），第二列为 Freebase 中对应电影实体的 ID，其示例片段如下图所示：

```
1291544   m.03177r
1291545   m.027pfg
1291546   m.01d1_s
1291550   m.053xlz
1291552   m.017jd9
1291554   m.01sxdy
```

所涉及的各种数据，包括实验一中所涉及的电影和书籍的 tag ，可在如下地址进行下载（密码：web2022）：https://rec.ustc.edu.cn/share/4a310a40-68d6-11ed-bcd0-45e65b70fc26

实验要求

本次实验要求分组完成，每组最多 3 人（可以少于 3 人，但无优惠政策）。

实验持续时间约为 4-5 教学周，实验报告的具体提交时间和更多详细要求将于第二阶段公布。

提交说明

请于截止日期（待定）前将实验二完整的实验报告（整个实验提交一份报告

即可）提交到课程邮箱 ustcweb2022@163.com，具体要求如下：

1. 邮件标题以及压缩包命名为"组长学号-组长姓名-实验 2-阶段 1"格式。邮件正文中请列出小组所有成员的姓名、学号。
2. 因未署名造成统计遗漏责任自行承担，你可以将邮件抄送你的队友。
3. 实验报告请务必独立完成，如果发现抄袭按 0 分处理。
4. 迟交实验将不被接收。
5. 后续版本会进一步更新具体实验报告要求。