实验 2 豆瓣电影的知识感知推荐

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主演）。

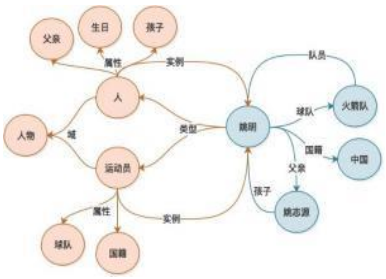
在本次实验中，我们要求各位同学从公开图谱中匹配指定电影对应的实体，并抽取合适的部分图谱，按照规则对抽取到的图谱进行处理（Stage1）；进而，基于豆瓣电影评分数据，结合Stage1所获得的图谱信息，进行可解释的、知识感知的个性化电影推荐（Stage2）。

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-TripleRDF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”



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

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1200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（Stage1）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。

[3]【选做】根据实验二提供的电影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。

[5]【选做】根据实验一中爬取的电影信息内容，考虑抽取文本中的实体和关系，加入到图谱中，增强电影本身的语义信息。

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。因为存在一类关系<http://rdf.freebase.com/ns/common.notable_for.display_name>，这类关系的构成三元组的尾实体通常为一种语言的字符串（如"ロマンティック・コメディ"@ja，表示日语的“浪漫喜剧”），而此类关系的尾实体一般不会和其他的实体相连。其他一些关系如<http://rdf.freebase.com/ns/people.person.date_of_birth>，作为存储数值类型的日期，也不会和其他的实体相连。约束实体前缀可以保证抽取的知识图谱更加精简，聚焦不同实体间的链接。

The object field may contain a Freebase MID for an object or a human-readable ID for schema from Freebase or other RDF vocabularies. It may also include literal values like strings, booleans and numeric values.

- ✧ 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。
- ✧ 对不同的Tag进行对齐或合并时，可以从不同角度考虑。如考虑不同Tag的共现关系，或利用模型（如word2vec）计算不同Tag在语义上的相似度。
- ✧ 在对电影信息内容中的实体和关系进行抽取时，需要预先定义抽取的实体和关系的范围。如针对剧情中的<演员，饰演，角色>, <角色，人物关系，角色>进行抽取，其中，人物关系可以进一步考虑约束为{对手，朋友，恋人}。然后可以借助一些模型如UIE(Universal Information Extraction，https://github.com/PaddlePaddle/PaddleNLP/tree/develop/model_zoo/uie)进行抽取。或者设计合适的Prompt，利用大模型（文心一言、讯飞星火、ChatGPT）进行抽取。
- ✧ **【选做】仅作为兴趣探索，不影响最终分数，时间不充裕情况下可以不做。**

数据集说明

我们提供了必要的数据文件，包括：

（1）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。

读取方法：

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

输出：

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

（2）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
1291555 m.01f8f7
1291557 m.01f85k
1291558 m.04j31dn
1291559 m.05_lhx
```

(3) 所涉及的其他数据，包括所涉及的电影的tag（Movie_tag.csv），电影的ID (Movie_id.csv)。

以上数据均可以从以下链接处下载。

链接：https://rec.ustc.edu.cn/share/16ed54c0-8751-11ee-b149-852f0d571a83

密码：web2023

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"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。

6.整个实验二只需提交一份实验报告，请等待第二阶段实验要求发布，并在全部完成实验二后再统一提交。