构建 FB15k 多模态数据集

https://github.com/JTWang722

由于 MMKG 多模态数据集中图片模态的数据存在着一些问题，如图片与实体不匹配、图片质量较差、图片链接无法打开等，我们尝试以 FB15k 数据集为基础，构建一个新的多模态数据集。我们尝试找到实体对应的 Wikipedia 链接，将 Wikipedia 页面中的图片作为视觉模态的数据。此外，我们还对图片进行了相似度的筛选，保证了实体与图片的匹配度。

一、将 FB15K 实体链接到现有数据库

由于 Freebase 服务关闭，无法获取 FB15k 实体的信息，现将 FB15k 数据集中的所有实体链接到现有的数据库，如 Wikidata、DBpedia 等，以便后续工作的开展。

1. 一部分实体链接至 Wikidata（14171 个）

Freebase API 关闭后，保留了一个包含 19 亿三元组的 dump 文件，除此之外，还提供了从 Freebase 到 Wikidata 的映射。利用这个映射关系文件，我们找到了大部分 FB15K 实体的 Wikidata 链接。

2. 一部分实体链接至 DBpedia（772 个）

剩余的一部分实体，我们将其链接至 DBpedia。

1）一部分利用 MMKB（https://github.com/nle-ml/mmkb）提供的 FB15k 与 DBpedia15k 之间的 sameAs 关系，找到了 438 个实体对应的 DBpedia 链接

2）剩余的 342 个实体，通过 sameAs 服务（http://sameas.org/）进行查询，该网站提供了不同数据库实体 URI 之间的等价关系。对于一个 FB15k 中的实体，如果查询到多个等价 URI，我们只保留第一个，通常是 DBpedia 链接。

3. 未找到任何链接（8 个）

最终剩余 8 个实体，通过上述方法均未找到其至现有数据库的链接。

二、获取 FB15k 实体的 Wikipedia 链接

将实体链接到现有的数据库后，使用其提供的服务接口可以很方便地进行查询。Wikidata 和 DBpedia 都提供了查询接口，利用其来获取实体对应的 Wikipedia 链接。最终，只有 53 个实体未找到 Wikipedia 链接，其余的 14898 个实体均映射至 Wikipedia。

1. 从 Wikidata 到 Wikipedia

利用 Wikidata 提供的 API 服务，根据 Wikidata ID 可直接查询其对应的 Wikipedia 链接。具体地，通过修改查询参数（实体的 Wikidata ID），获得查询结果页面 URL，通过网页爬虫，得到实体对应的英文 Wikipedia 链接。

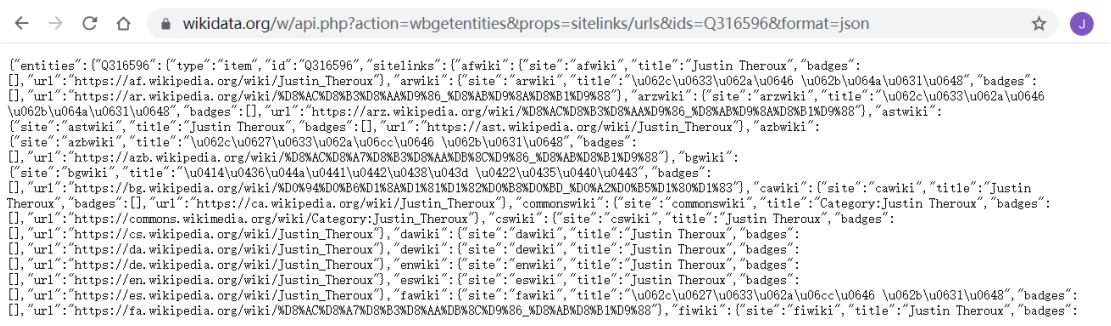

图 1 ID=Q316596 实体的查询结果页面

2. 从 DBpedia 到 Wikipedia

利用 DBpedia 提供的 SPARQL 服务，通过实体的 foaf:isPrimaryTopicOf 属性找到其对应的 Wikipedia 链接。具体地，通过修改 SPARQL 语句中的实体 URL，获取每个实体的 foaf:isPrimaryTopicOf 属性值，爬取查询页面中的 Wikipedia 链接。

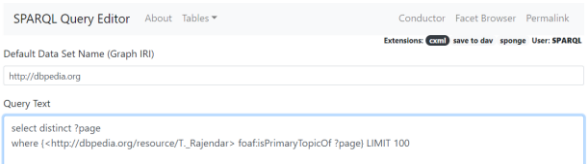

图 2 <http://dbpedia.org/resource/T._Rajendar>实体的 SPARQL 查询页面

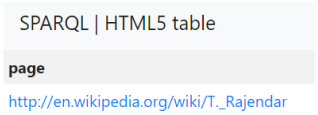

图 3 <http://dbpedia.org/resource/T._Rajendar>实体查询结果

三、爬取 FB15K 实体对应的 Wikipedia 页面中的图片

在得到实体对应的 Wikipedia 页面后，利用爬虫，获取页面中的所有图片。我们选取了 class 属性值为 infobox（infobox vcard、infobox biography vcard、infobox geography vcard、infobox vcard plainlist）、thumbinner 以及 gallery mw-gallery-packed 的标签，爬取这三类标签下所有图片的 URL。

获取到每个实体对应的图片 URL 后，为进行下一步的筛选，需要将其下载下来。在这个过程中，有误的图片 URL 将被过滤。

四、通过衡量图片之间的相似度进行筛选

接下来，我们需要对每个实体对应的图片集进行筛选。我们的思想是相同实体对应的图片应该是相似的。每个实体至多保留 5 张图片。

1. 计算图片特征向量

利用预训练的 Resnet50 与 Vision Transformer 模型，计算每个实体每张图片的特征向量。

2. 根据相似度进行筛选

对于一个实体的图片集，计算每张图片与剩余图片的余弦相似度之和，将两个模型的结果加权求和，保留值最高的 5 张图片作为该实体最终的图片集。

五、FB15k 多模态数据集

1. 结构

每个实体对应一个文件夹，文件夹名为实体的编号，文件夹内是实体对应的图片，若为空则该实体无图片。

2. 图片数量统计

该数据集总共有 14191 个实体，总共有 48104 张图片。

图片数量	0	1	2	3	4	5
实体个数	1186	2850	1661	1485	1368	6401

六、改进

1. 缺乏评价指标，对数据集进行评价；
2. 计算图片特征向量时，只计算了三通道彩色图片，没有计算灰度图。对于灰度图，可利用 convert(‘RGB’)转换后再输入网络进行计算；
3. 图片分辨率较低，尺寸较小；
4. 图片数量不充分（大部分实体图片数量小于 5）。