KDI : Knowledge and Data Integration

‘KDI-movie’

Yuhao Mu,Kailong Suo,Pengyu Huang

Contributors

- Data Scientist: Yuhao Mu
- Domain Expert: Yuhao Mu
- Knowledge Engineer: Yuhao Mu
- Knowledge Engineer: Kailong Suo
- Project Manager: Pengyu Huang
- Tutor Data: Rui Zhang
- Tutor Knowledge: Rui zhang
- Speaker: Yuhao Mu

Table of Contents

Project description

As one of the important forms of human entertainment and cultural expression, movies carry rich artistic, cultural and entertainment values. The knowledge graph of movie information aims to organize and present movie-related entities, attributes, and relationships to help users quickly obtain accurate and comprehensive movie-related information.

The project aims to provide a unified, comprehensive and structured solution for movie information-related queries based on the knowledge graph method by cleaning, organizing and integrating official and authoritative data on the Internet. This solution can effectively meet the movie needs input by users (such as movie genres, actors, etc.) and help users quickly obtain accurate and comprehensive movie information.

Initial definitions

- The extracted CQs can be mainly categorized into three types: queries related to actors, queries related to movies, and queries related to movie genres.
- We have extracted three objects, namely movies, actors, and movie genres.
- Although movie genres may seem like a property of movies, due to their many-to-many relationship with movies, we have chosen to treat movie genres as a separate concept.

Data resources

- Data is obtained from "The Movie Database (TMDb)", the website address is https://www.themoviedb.org. The official provides registered user API KEY for querying and downloading data.
- The data acquisition method in this example: use Stephen Chow as the initial entry to obtain all the movies in which he has appeared; then obtain all the actors who have participated in these movies; and finally obtain all the movies in which all the actors have appeared.

Data resources

- The actor's basic information includes: name, English name, date of birth, date of death, place of birth, and personal profile. The basic information of the movie includes: movie name, movie introduction, movie rating, movie release date, and movie type
- We narrow the data set by calling the website's search interface to query specific data, and use Python scripts to sample and clean the website data to make it structured; we rationally discard the characteristics of the result data through sampling and cleaning under different search restrictions. , build a heterogeneous database group, which is the data source of the project.

Knowledge resources

- Our information schema primarily includes three classes: movies, genres, and actors. It encompasses the relationships of an actor appearing in a particular movie, the relationship of a movie containing certain actors, and the types of genres associated with a movie. Additionally, there are three Object Properties: "actors in a movie," "movies containing certain actors," and "genres of a movie." Each class also has its own set of data properties.
- If the correspondence between movies and actors is set as one, queries in one direction may be simple, but in the other direction, they could be challenging. However, if set as two, it would consume more time during information retrieval, and managing the crawled information would also become more complex.

Knowledge resources

- In the end, we chose to define two relationships. During information crawling, we only retrieve information about which movies an actor has appeared in. Subsequently, we utilize Apache Jena's reasoning capabilities to automatically derive the complementary relationship, obtaining information about which actors are associated with a particular movie.
- We use Protege to formalize the proposed schema.

Metadata collection

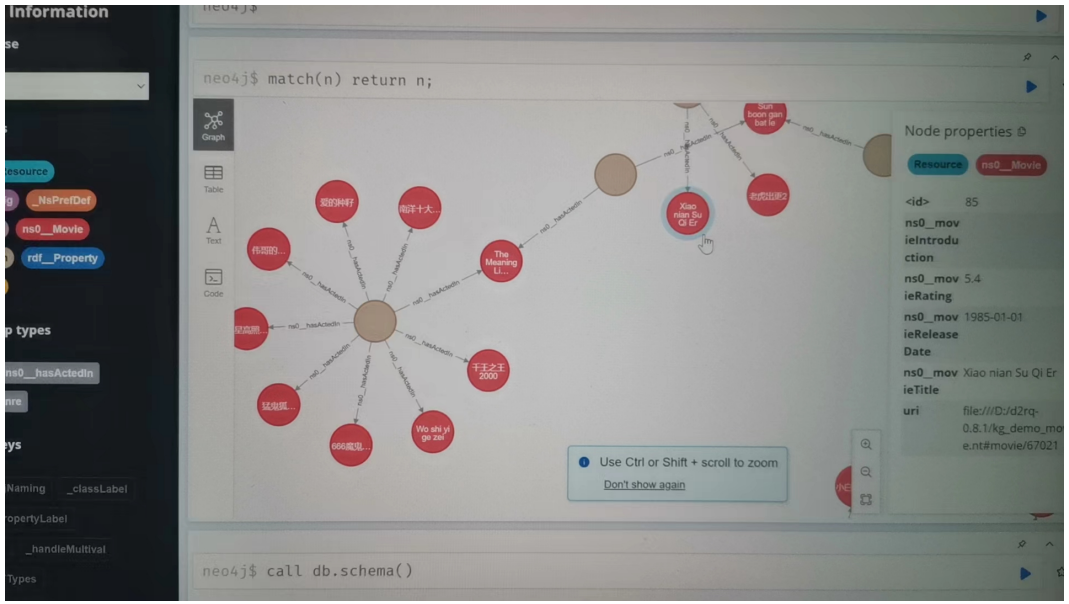
- We assigned ID numbers to actors, movies, and categories.
- We noticed that these ID numbers for actors, movies, and categories were missing in the original data, so we added them when we found out.
- We created a feature that shows the types of movies an actor has been a part of.
- This feature allows us to find out which movie genres a specific actor has worked in.
- The tricky part was that actors and genres weren't directly connected. We had to go through the movies to indirectly get that information.

Problems and Solutions

- In the process of data collection, when encountering issues of low data quality and inconsistency, we adopt data cleaning methods to judiciously discard unnecessary data.
- We are not sufficiently proficient in the knowledge and data fusion aspect of the project. While working on the project, we are also learning through research or seeking advice from others
- Initially, communication and collaboration issues were affecting the progress of the project. Ultimately, we decided to have regular discussions, which effectively resolved this problem.
- The relationships of which actors participated in a movie and what types of movies an actor has played in cannot be directly obtained.We finally decide to use Apache Jena's reasoning capabilities to retrieve these two types of information.

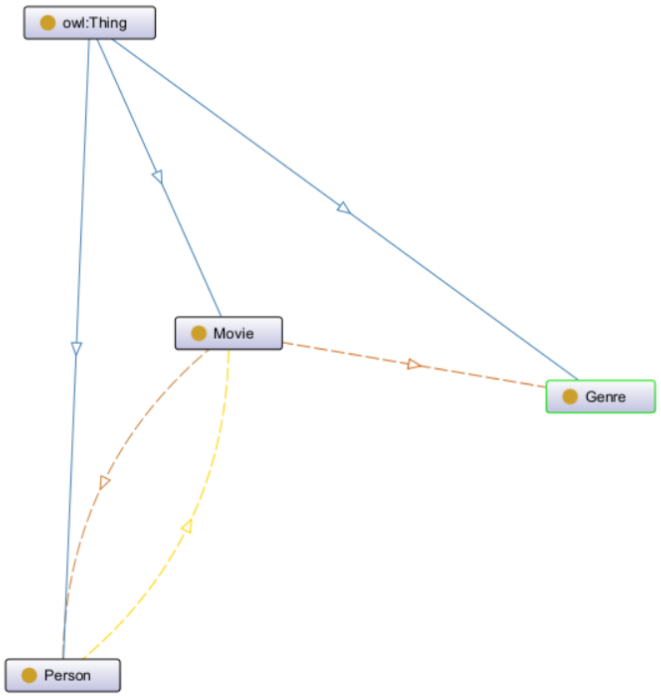
Outcome and Analytics

- Our Knowledge Graph ultimately achieved the capability to perform queries such as retrieving actors based on a movie, finding movies where two actors collaborated, and determining the types of movies a particular actor has played in.



Outcome and Analytics

Ontology



Outcome and Analytics

Instance

```
1 for input a movie name to search who act in
2 for input an actor name to search his or her movie
3 for input two actor names to search which movie they act together
4 for input an actor name to search the genres of movies he or she has acted in
5 for input an actor name and a genre to search the movies which has the input genre and he or she has acted in
1
功夫
x
元华
林雪
元秋
黄圣依
冯小刚
周星驰
陈国坤
```

```
0 for exit
1 for input a movie name to search who act in
2 for input an actor name to search his or her movie
3 for input two actor names to search which movie they act together
4 for input an actor name to search the genres of movies he or she has acted in
5 for input an actor name and a genre to search the movies which has the input genre and he or she has acted in
2
元秋
x
功夫
Ding tian li di
女警察
Mei nui sik sung
Wu zhao shi ba fan
越光宝盒
```

Outcome and Analytics

Instance

```
0 for exit
1 for input a movie name to search who act in
2 for input an actor name to search his or her movie
3 for input two actor names to search which movie they act together
4 for input an actor name to search the genres of movies he or she has acted in
5 for input an actor name and a genre to search the movies which has the input genre and he or she has acted in
3
元华
林雪
x
功夫
野蛮密笈
道士下山
please input the number of question
```

```
0 for exit
1 for input a movie name to search who act in
2 for input an actor name to search his or her movie
3 for input two actor names to search which movie they act together
4 for input an actor name to search the genres of movies he or she has acted in
5 for input an actor name and a genre to search the movies which has the input genre and he or she has acted in
4
元华
x
冒险
剧情
爱情
动作
惊悚
犯罪
奇幻
喜剧
纪录
科幻
恐怖
家庭
战争
```

Outcome and Analytics

Instance

```
0 for exit
1 for input a movie name to search who act in
2 for input an actor name to search his or her movie
3 for input two actor names to search which movie they act together
4 for input an actor name to search the genres of movies he or she has acted in
5 for input an actor name and a genre to search the movies which has the input genre and he or she has acted in
5
元华
喜剧
x
功夫
福星高照
龙的心
Fei lung mang jeung
Hung kuen dai see
越光宝盒
鬼猛脚
Tou shen gu zu
急冻奇侠
野蛮密笈
Xun zhao Cheng Long
雀圣
龙的传人
```

Open issues

This final section aims to describe issues eventually remained open developing the project. More in detail try to answer the following questions:

- In the end, we successfully represented all the information outlined in the initial definition within our knowledge graph.
- The main challenge lies in the information about what types of movies an actor has played in, as this information does not directly exist in the original data.
- Except for actors, other movie crew members have not been included in the Knowledge Graph because their numbers are vast, and they are not deemed necessary for solving the problems in the given application scenarios.
- Our Knowledge Graph still has some issues, such as the omission of information about the directors of movies and their details.

KDI : Knowledge and Data Integration

‘KDI-movie’

Yuhao Mu,Kailong Suo,Pengyu Huang