A content-based movie recommendation engine

Ana Carolina Camargos Couto

Department of Applied Mathematics
Western University
London, ON

Problem Statement

- Problem: given an active user and their rating history, recommend movies they would probably like to watch.

- Motivation: strengthen the quality of the service offered by the platform and increase the amount of sales/revenue.

Data Overview

‘The Movies Dataset’ from kaggle.com

Movie table (45,000 movies):

movieID, budget, popularity, revenue, runtime, language, genres, release_date, title, overview, tagline

User table (270,000 users):

userID, movieID, rating

Data Investigation

- Structure: in the movie table, the data exhibits a variety of types: text, dictionary, datetime, and numerical with different ranges.

- Information: the richness of metadata describing each movie’s contents and attributes makes the dataset suitable for content-based models.

- From the user table, we have the rating history of each user

The Methodology

- For each user, one model. The training data for each user is obtained by joining the user table and the movie table on the movieID key.

Training data: [userID, movieID, rating, [movie_attributes]]

- 15% of the training data of each user is held-out for testing purposes. This consists of a few previously rated movies.

- A Support Vector Regression model is fit to the movie attributes in the training data, with the ratings for labels. The model can then be used to predict the ratings in the test set, as well as the ratings of not-yet-seen movies.

Data Preprocessing

Movie table:

1. Replacement of NaN values in the numerical columns with the column average, and normalisation of values to [0,1]

2. Extraction of relevant keys from dictionary attributes and conversion to one-hot-encoding.

Ex: [{'id': 12, 'name': 'Adventure'}, {'id': 14, 'name': 'Fantasy'}, {'id': 10751, 'name': ‘ Family’}] becomes [0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Data Preprocessing

3. Tfid vectorisation of sentences (title, overview, tagline):

Ex: 'Two identical strangers. Two different worlds. One perfect match.'
becomes [0,0,0,…,0.36902202,…,0.92942065,…,0,0,0] with 200 components total

4. Dimensionality Reduction:

The movie table after preprocessing has 718 features. In order to project the data in a lower dimensional space, Principal Component Analysis was applied. The data size was reduced to 2 features, with a PCA variance of 0.9996.

Results

User 4:

```
In [27]: recommendations(4)
['Ali', ['Drama'], 'The Chronicles of Riddick', ['Action
ScienceFiction'], '10,000 BC', ['Adventure Action Drama
Fantasy'], 'The Martian', ['Drama Adventure
ScienceFiction'], 'Live by Night', ['Crime Drama']]
Out[27]: 2.059849480556316
```

User 123:

```
In [27]: recommendations(123)
['Dinosaur', ['Animation Family'], "Harry Potter and the
Philosopher's Stone", ['Adventure Fantasy Family'], 'The
Island', ['Action Thriller ScienceFiction Adventure'], 'The
Hunger Games: Mockingjay – Part 1', ['ScienceFiction
Adventure Thriller'], 'Night at the Museum: Secret of the
Tomb', ['Adventure Comedy Fantasy Family']]
Out[33]: 1.0190024473936012
```

Results

User 2860:

```
In [36]: recommendations(2860)
['Planet of the Apes', ['ScienceFiction Adventure Drama
Action'], 'Hombre', ['Action Drama Western'], 'Walker',
['Drama History Western'], 'Crossover', ['Action Adventure
Drama'], 'Burning Palms', ['Comedy Drama Family']]
Out[36]: 0.9945076966246945
```

User 31864:

```
In [38]: recommendations(31864)
['The Polar Express', ['Adventure Animation Family
Fantasy'], 'How to Train Your Dragon', ['Fantasy Adventure
Animation Family'], 'Shrek Forever After', ['Comedy
Adventure Fantasy Animation Family'], 'Wreck-It Ralph',
['Family Animation Comedy Adventure'], 'Big Hero 6',
['Adventure Family Animation Action Comedy']]
Out[38]: 0.9622876432478998
```

Validation

- Accuracy: the RMSE of each user model tells us how far off we are from the possible rating.

- Novelty: the number of different genres recommended to a user indicates whether the generated suggestions are diverse.

Questions?

Thank you.