

THE CHINESE UNIVERSITY OF HONG KONG, SHENZHEN

CSC 3170

DATABASE SYSTEM

---

## Group 26 Report: Rotten Potatoes

---

*Authors:*

郑时飞  
朱伯源  
李易  
施天昊  
汪明杰

*Student Number:*

119010465  
119010485  
119010156  
120090472  
119010300

Thursday 12<sup>th</sup> May, 2022

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Design</b>	<b>2</b>
2.1	Entity-Relationship Model . . . . .	2
2.2	Relational Schema . . . . .	3
2.3	Constraint . . . . .	4
2.4	Index . . . . .	5
<b>3</b>	<b>Implementation</b>	<b>5</b>
3.1	Frontend . . . . .	5
3.2	Backend . . . . .	6
3.3	Web Crawler . . . . .	6
3.3.1	Basic Workflow . . . . .	6
3.3.2	Encountered Problems and Solutions . . . . .	7
3.4	Sample Queries . . . . .	8
3.4.1	User related . . . . .	8
3.4.2	CRUD on movies, actors and directors . . . . .	8
3.4.3	Comments related . . . . .	8
3.5	Data Analysis . . . . .	8
3.5.1	Recommendation System with Restricted Boltzmann Machines . . . . .	8
3.5.2	Neighbourhood Model . . . . .	9
<b>4</b>	<b>Future Works</b>	<b>10</b>
<b>5</b>	<b>Result</b>	<b>10</b>
<b>6</b>	<b>Conclusion</b>	<b>10</b>
<b>7</b>	<b>Self Evaluation</b>	<b>10</b>
7.1	What we have achieved . . . . .	10
7.2	Future Improvements . . . . .	10
<b>8</b>	<b>Contribution</b>	<b>11</b>

# 1 Introduction

As an art and commercial work, movies have become pursuits of many people and generated tremendous value. Currently, there are two types of online movie platforms. On the one hand, platforms including Rotten Tomatoes and Douban gives people rich information regarding movies. On the other hand, sites like Netflix offers precise, personalized recommendation to users. Such a recommendation strategy allows users to discover more films they enjoy and attracts more users to the platform.

However, currently, no site combines the strengths of the mentioned movie sites. Besides, existing popular movie websites are full of quarrels and controversies from the perspective of the community environment. The movie recommendation sare influenced deeply by the advertising, not the quality of films. Therefore, to satisfy the real requirements of movie lovers, we build a movie database called **Rotten Potatoes**, based on which, users can talk about their reviews on the movie freely. It also offers plenty of searching functions and personalized recommendations from the platform.

We use python packages *Requests*, *BeautifulSoup* to obtain the required data from IMDB, store it in a MySQL database. Then we build our searching functions and personalized moive recommendations on those data. We also build a user-friendly web using *amis* as the front end and *express* as the back end.

## 2 Design

In this section, we focus on the design of Entity-Relationship Model, the reduction from ER diagram into relational schemas, constraint and index.

### 2.1 Entity-Relationship Model

As shown in the [ER digram below](#), there are 6 entities and 5 relationships.

**Entity “movies”** To store id, name, cover url, introduction, release year and genres of a movie, where genres is a multivariate attribute. Identified by id.

**Entity “directors”** To store id, name, photo url, introduction and birth date of a director. This entity has a one-to-many relationship “direct” with the entity “movies”, which means a director can directs multiple movies and a movie can be directed by only one director in our assumption. Identified by id.

**Entity “actors”** To store id, name, photo url, introduction and birth date of an actor. This entity has a many-to-many relationship with the entity “movies”, which means an actor can act in multiple movies and a movie can be acted by multiple actors in our assumption. Identified by id.

**Entity “users”** To store id, name, avatar url, password of a user. This entity has a many-to-many relationship with the entity “movies”, which means a user can comment on multiple movies and a movie can be commented by multiple users in our assumption. Identified by id.

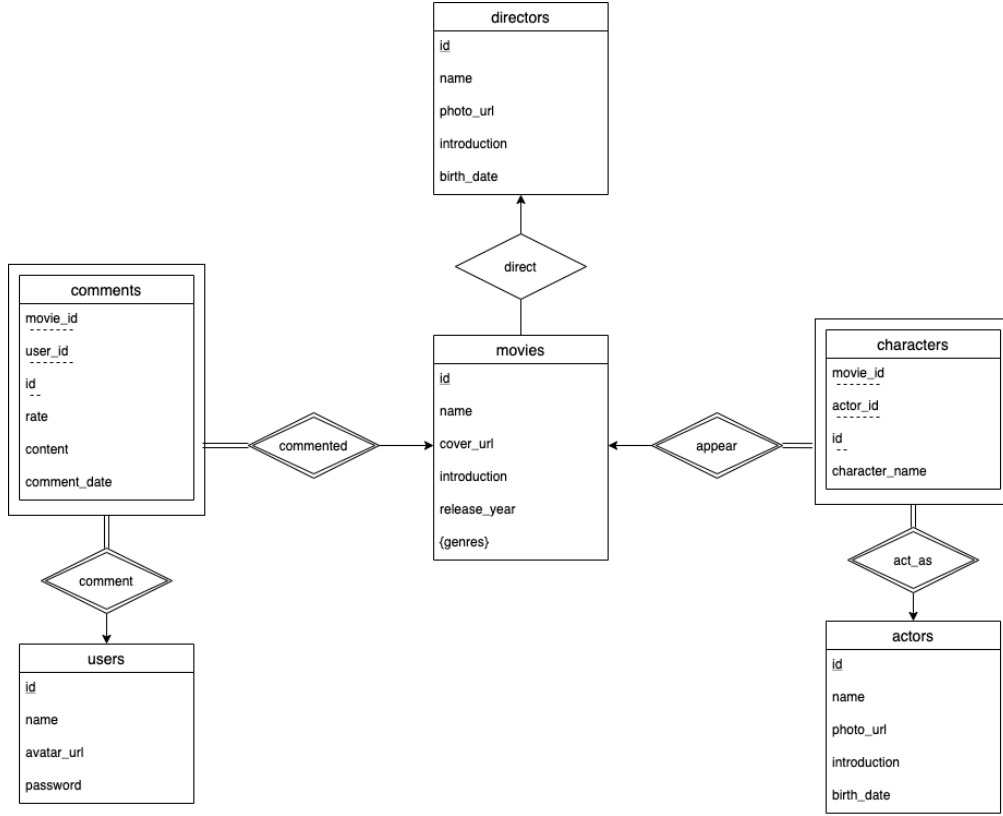


Figure 1: ER Diagram

**Entity “characters”** To store movie id (of the movie where this character appears), actor id (of the actor who acts as this character), id, character name of a character. This entity is a weak entity identified by entity “movies” through relationship “appear” and entity “actors” through relationship “act as”, and also by its own id, which means a character can be acted by exactly one actor and appear in exactly one movie in our assumption (we treat characters of the same name appearing in multiple movies or acted by multiple actors as multiple different characters for simplicity). Note that since this entity is also identified by its own id, it is allowed that an actor acts as multiple characters in the same movie.

**Entity “comments”** To store movie id (of the movie commented by this comment), user id (of the user who makes this comment), id, rate (from 0 to 10), content and comment date of a comment. This entity is a weak entity identified by entity “movies” and entity “users”, and also by its own id, which means a comment is on exactly one movie and is made by exactly one user in our assumption. Note that since this entity is also identified by its own id, it is allowed that a user makes multiple comments on the same movie.

## 2.2 Relational Schema

As shown in the [relational schema digram below](#), there are 7 schemas.

The following reductions are made:

- The attribute “genres” in entity “movies” is reduced to schema “genres” with attributes “genres\_name” and “movie\_id” as a foreign key, both of which forms a primary key to make

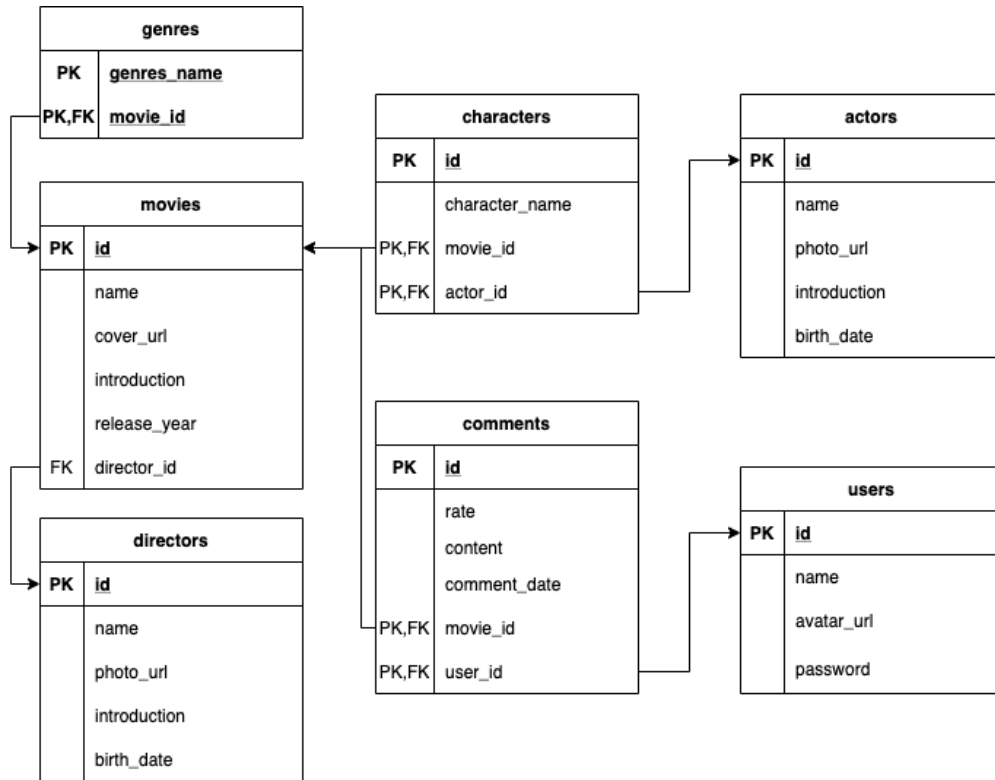


Figure 2: Relational Schema Diagram

sure no redundant genres of a movie.

- The relationship “direct” between entity “movies” and “directors” is reduced to attribute “director\_id” as a foreign key in schema “movies” so that a movie is directed by exactly one director.
- All attributes identifying the entity are reduced to primary keys.
- “movie\_id”, “actor\_id” of entity “characters”, “movie\_id”, “user\_id” of entity “comments” are reduced to foreign keys in their schemas referencing their corresponding identifying strong entities.

## 2.3 Constraint

3 types of constraints are further added:

### Not Null Constraints

- All “id” attributes as they are primary key and thus automatically becoming not null.
- “name” attribute of schema “movies”.
- “name” and “director\_id” attributes of schema “directors”, as a movie is directed by exactly one director.
- “name” attribute of schema “actors”.

- “name” and “password” attributes of schema “users”.
- “actor\_id”, “movie\_id” and “character\_name” attributes of schema “characters”, as it is a weak entity of schemas “actors” and “movies”.
- “user\_id”, “movie\_id”, “rate”, “content” and “comment\_date” attributes of schema “comments”, as it is a weak entity of schemas “users” and “movies”.
- “genres\_name”, “movie\_id” attributes of schema “genres”, as it is a multivariate attribute of schema “movies”.

**Unique Constraint** A unique constraint is added to “name” attribute of schema “users” as by our assumption there should be no repeating user names.

**Check Constraint** A check constraint is added to “rate” attribute of schema “comments” to make sure the rate is from 0 to 10.

## 2.4 Index

The following attributes are indexed to make search faster:

- “name” and “release\_year” attributes of schema “movies”.
- “name” and “birth\_date” attributes of schema “directors”.
- “name” and “birth\_date” attributes of schema “actors”.
- “name” attribute of schema “users”.

# 3 Implementation

## 3.1 Frontend

The frontend of our project is constructed using AMIS, a low-code front end framework. It can generate a website using json configurations, which is suitable for developing a light-weight and agile application like this project.

When users access the webpage, they will be directed to the Login page, where registered user can input user name and password to log in. We also added support for new user registration. After login, users can see the main structure of our website, which features a navigation pannel on the left side and the content on the right. We constructed the following pages:

- Movie: a list of all the movies with their name and posters. Each movie is a ”Card” widget linking to the movie detail page.
- Actor: a list of all the actor/actresses, also implemented using the ”Card” widget, containing links to the detailed information.
- Director: a list of all the directors implemented similarly as the **Actor** page
- Comment: this page contains the latest comments that users release.
- Me: a portal for users to edit their information. Including updating avatar, name, password. This page also contains a list of movies recommended to the user.

- Search: to search for movies/actors/user, we implemented three different pages. The details will be discussed in **Sample Queries** section.

Fig.3 shows the movie page of our website. Since our application is user-oriented, the UI of our website is very concise and user-friendly.

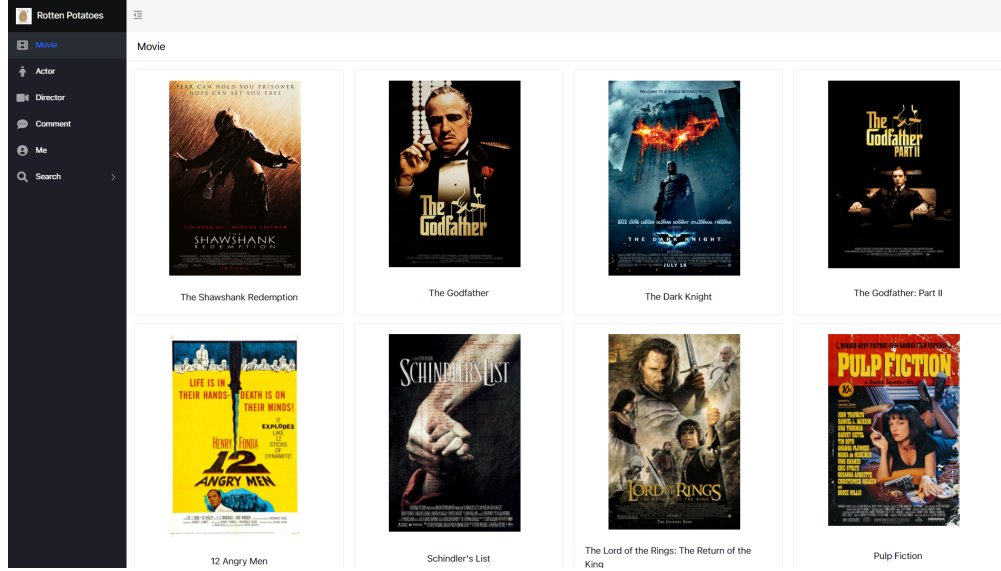


Figure 3: Movie page of the website

## 3.2 Backend

Backend of this project is implemented through the express package of javascript. During the initialization, the 'init.js' will read data from CSV files and create queries to batch-insert records into corresponding tables.

## 3.3 Web Crawler

### 3.3.1 Basic Workflow

To populate our database with real-world data, we wrote a web crawler to scrap information from IMDB's Top 250 Movie Chart. The crawler is written in Python. It uses the requests library to send https requests. The desired fields are acquired through parsing the page using BeautifulSoup 4 library. The detailed process is as follows (see Fig.4):

1. Access the Top 250 Chart, where all the URLs to the detailed movie pages are located
2. Traverse through the chart. For every movie of the chart:
  - (a) Access the detailed movie page, where information regarding the movie can be found.
  - (b) From the detailed movie page, we access the casting information.
 

For the director:

    - For directors yet to be recorded, access the detailed director page, where information regarding the director can be found.

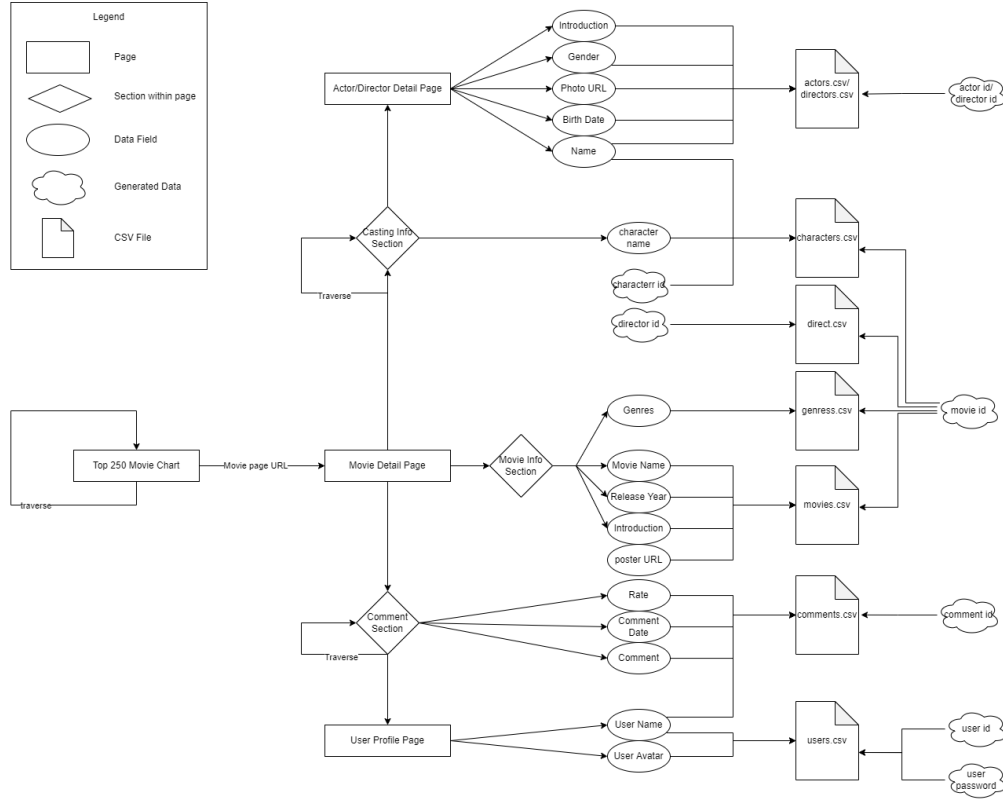


Figure 4: Web crawler flowchart

For the actors/actresses who casted in this movie:

- Access the detailed actor/actress page and find detailed information.
- (c) The comment section locates at the bottom of the movie page. We collect the rating, comments from this section.
- (d) For each comment, we acquire the user information by accessing their homepage.

### 3.3.2 Encountered Problems and Solutions

During the scrapping process, we encountered some issues.

- The same director or actor/actress can participate in different movies. Therefore, we keep a mapping relationship and check wheter we have record the same person's information. This step is done through hashing the person's name, which takes  $O(1)$  complexity.
- We do not have access to each user's password, so we randomly generate a password for each user we scrapped from IMDB. The password is a random combination 5 to 10 numbers and characters.
- Some actors' birth date cannot be achieved from their detailed page. We leave them as NULL.
- The gender of the movie stars is not explicitly listed on their detail page. However, by seeking their role in the movie (i.e, actor/actress), we can acquire their gender.



### 3.4 Sample Queries

#### 3.4.1 User related

#### 3.4.2 CRUD on movies, actors and directors

#### 3.4.3 Comments related

### 3.5 Data Analysis

#### 3.5.1 Recommendation System with Restricted Boltzmann Machines

In 2007, Hinton showed that Restricted Boltzmann Machines (RBMs) can be successfully applied to Netflix dataset and do personalized movie recommendations (Hinton et al., 2007). This paper was given in a DDA course project, but that project requires only the basic RBM without biases, conditional, neighborhood. We reproduce the model proposed by Hinton (see Fig.5) in this project to do recommendation. We use Contrastive Divergence (CD) to approximate the maximum likelihood of the parameters by Gibbs sampling.

- First get the visible movies ratings  $\mathbf{V}$  of the user.
- Sample the hidden units to binary numbers by probability given by  $p(h_j = 1 | \mathbf{V}, \mathbf{r}) = \sigma(c_j + \sum_{i=1}^m \sum_{k=1}^K v_i^k W_{ij}^k + \sum_{i=1}^M r_i D_{ij})$ , where  $\sigma(x) = \frac{1}{1+e^{-x}}$ .
- Reconstruct the visible units by  $p(v_i^k = 1 | \mathbf{h}) = \text{softmax}(b_i^k + \sum_{j=1}^F h_j W_{ij}^k)$ , where  $\text{softmax}(x_k) = \frac{e^{x_k}}{\sum_{l=1}^K e^{x_l}}$ . Notice that we only reconstruct the movies the use has rated.
- Update the parameters by the origin data and reconstructed data. For example,  $\Delta W_{ij}^k = \epsilon(<v_i^k h_j >_{data} - <v_i^k h_j >_T)$ , where  $T$  represents the number of runs we do Gibbs sampling.
- Prediction is much the same as doing Gibbs sampling except that we can reconstruct (predict) missing ratings.

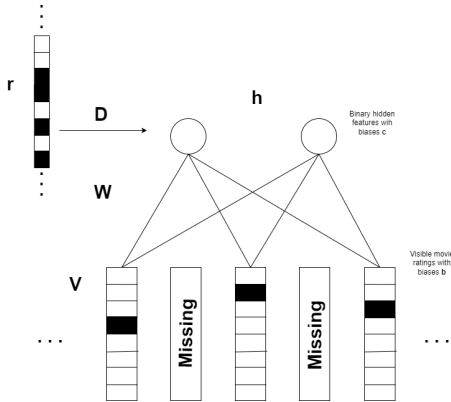


Figure 5: Conditional RBM

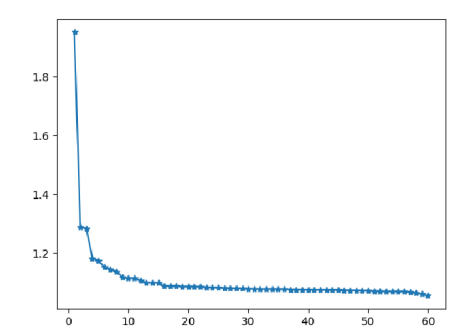


Figure 6: Train RMSE vs. Epochs

Fig.6 shows the evolution of Root Mean Squared Error during training. We achieve a RMSE of 1.04. Every time the user enters **Me** page on our website, the system will automatically predict the ratings over all movies and recommend the top 8 movies to the user in **Guess you like** module. See Fig.7 for an example.

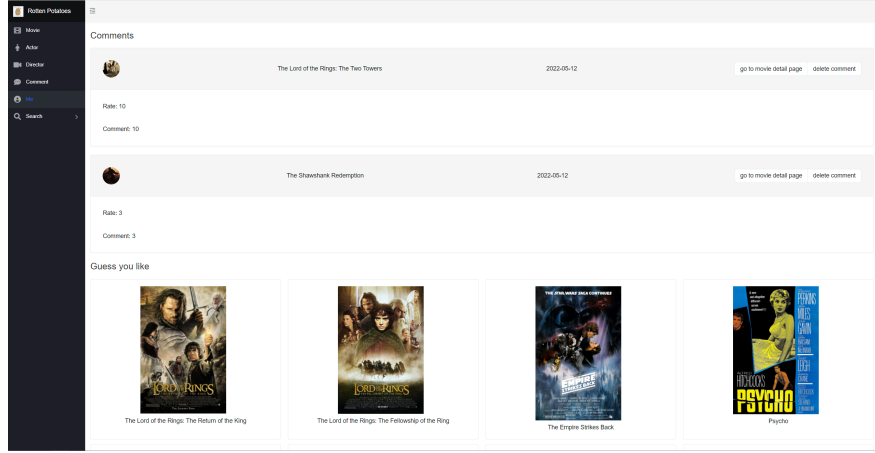


Figure 7: Personalized Movie Recommendation

### 3.5.2 Neighbourhood Model

Users have their preferences and tend to give similar ratings to movies of the same kind. In other words, movies are correlated and we can utilize this property to improve our predictions  $\hat{R}$  :

- Define an error matrix by  $\tilde{R} = R - \hat{R}$
- Define similarity between two movies by their column features (user ratings):  $d_{AB} = \frac{\tilde{r}_A^T \tilde{r}_B}{\|\tilde{r}_A\|_2 \|\tilde{r}_B\|_2}$
- Predict the error of one movie through the errors of its neighbours:  $\hat{r}_{iA} = \frac{1}{\sum_{B \in S} |d_{AB}|} \sum_{B \in S} d_{AB} \tilde{r}_{iB}$ , where  $S$  is the set of neighbours (with high absolute similarities) of  $A$ .
- Improve our RBM predictions by  $R_{um}^* = \hat{R}_{um} + \hat{r}_{um}$  for each  $(user, movie)$  pair desired.

We improve the RMSE from 1.04 to 1.02. The improvement is not high because our matrix is sparse. In our data set, many users only give one rating, so it is hard to construct a well defined similarity matrix. The other reason is that we crawl high-rating movies and the biases are not large to give a performance boost. However, we are still able to leverage the data and get some insight into the movie ratings.

For the movie *The Godfather*, movie *The Godfather: Part II* has a high similarity (based on the difference between real rating and predicted rating) of 0.9999, which is expected. However, movie *The Lord of the Rings: The Return of the King* has a similarity of  $-0.9999$  while the movie *The Lord of the Rings: The Fellowship of the Ring* has a similarity of 0.9999. This is counter intuitive and we should look into the data. We have three users giving ratings to both *The Godfather* and *The Lord of the Rings: The Return of the King*. The ratings are (10,10), (10,10) and (8,10), respectively. For *The Godfather* and *The Lord of the Rings: The Fellowship of the Ring*, the rating pairs are all (10,10). 8 is a relatively low rating in our dataset, hence the algorithm concludes that they are high negatively correlated.

## 4 Future Works

## 5 Result

## 6 Conclusion

In this project we build a movie database based on a large amount of data collected from **IMDB**. Additional attention is paid to assure a **BCNF** form is kept for every entity in our database, thus the redundancy part is eliminated. We also **implement useful functions** for our users to search for movies, casts or other users using name, time information, rate(only movies) and genres(only movies) as the keyword.

In order to improve user experience and replace the complex database operations by a few simple mouse clicks from users, we build a user friendly website. Based on that, abundant **personalized interaction behaviors** can be achieved, which serves our users more conveniently. We hold an integral process of account management for users. Users can add comments to a movie and rate on it as well as discover their appreciated reviews or users.

It is common for a movie database to recommend movies that the users have potential interest in. However, how to make a personalized recommendation lists of high quality is always a problem. We use a lot of techniques in the data analysis to satisfy the needs from users . We build a **Restricted Boltzmann Machines** with **Neighbourhood model** and get gratifying results.

## 7 Self Evaluation

In this part, we evaluate the advantages and disadvantages of our project/

### 7.1 What we have achieved

1. Build a database conforming to the low redundancy paradigm
2. Implement and optimize many necessary searching functions
3. Create a user-friendly website to lower the bar of use
4. Feed the database with a suitable amount of data and design a proper data analysis algorithm to satisfy personalized requirements

### 7.2 Future Improvements

1. More information about movies, directors, actors can be added to the database. For example, for actors we may add their nationality and Oscar nominations; for movies we may add its budget and the classification (PG13, NC17).
2. More interaction functions can be add, like “like” a movie or “subscribe” a user. We can also add functions like “want to watch” a movie or “watched” a movie without commenting on it. This information can be utilized in conditional RBM to provide better recommendation (Hinton et al, 2007). Currently, we have the algorithm implemented but the information is unknown.

3. In the future, we plan to distinguish the users of our platform. We will allow a common user to upgrade as administrator, which will reduce the effort to maintain and supervise the comments other users post.
4. Currently the web crawler is implemented using synced requests, which reduced its performance. In the future, we will write the crawler using async requests to speed up the crawling performance and acquire more data for our database.

## **8 Contribution**