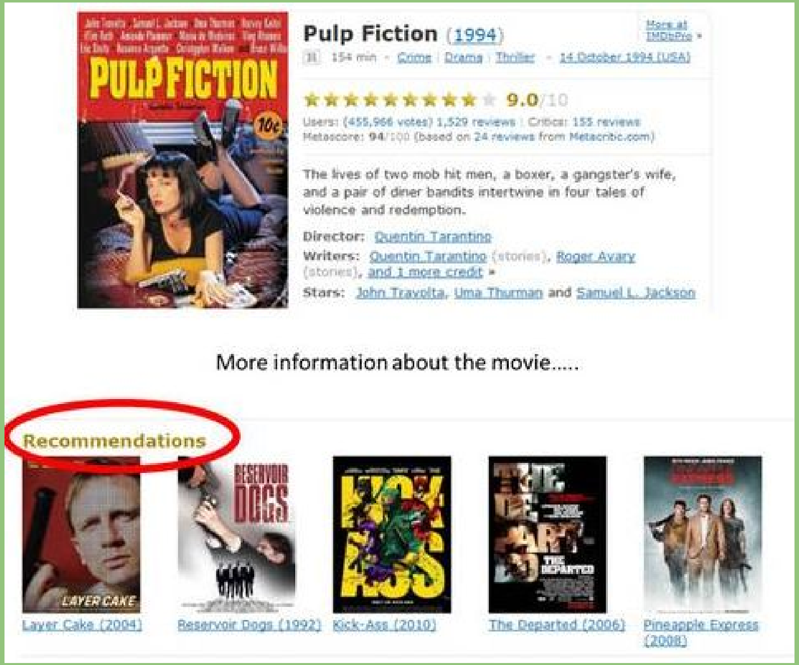
Clustering Based Recommendation Systems

Capstone Project: Team #14

1. Rahul
2. Siddhant
3. Himangshu
4. Sushma
5. Krunal
6. Harsha

Introduction

INTRODUCTION



Pulp Fiction (1994)

R 154 min - Crime | Drama | Thriller - 14 October 1994 (USA)

More at IMDbPro »

★★★★★★★★★☆ 9.0/10

Users: (455,966 votes) 1,529 reviews | Critics: 155 reviews
Metascore: 94/100 (based on 24 reviews from Metacritic.com)

The lives of two mob hit men, a boxer, a gangster's wife, and a pair of diner bandits intertwine in four tales of violence and redemption.

Director: Quentin Tarantino
Writers: Quentin Tarantino (stories), Roger Avary (stories), and 1 more credit »
Stars: John Travolta, Uma Thurman and Samuel L. Jackson

More information about the movie.....

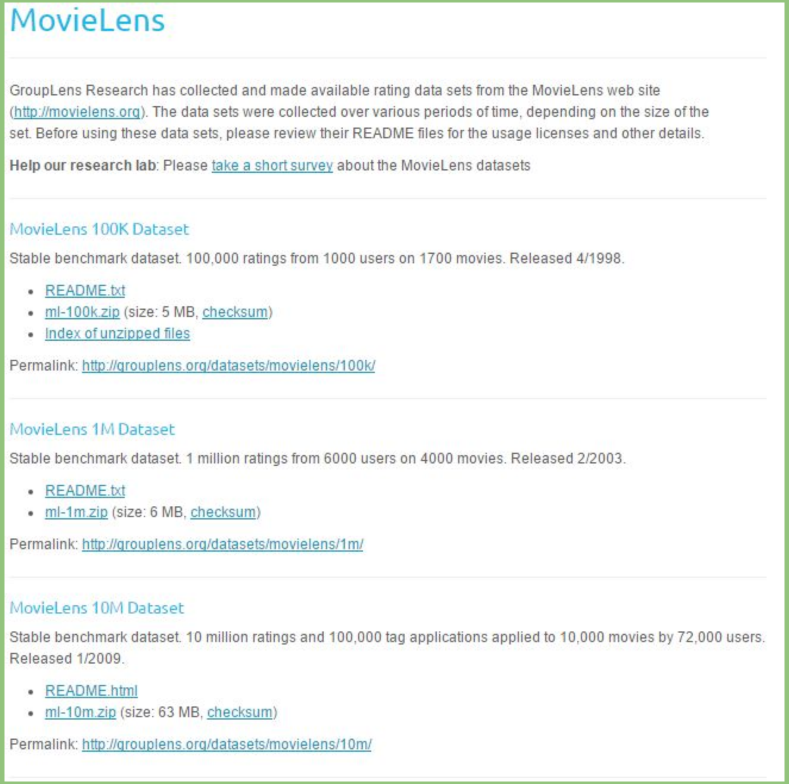
Recommendations

Layer Cake (2004)
Reservoir Dogs (1992)
Kick-Ass (2010)
The Departed (2006)
Pineapple Express (2008)

- Recommendations systems give two kinds of feedbacks:
    - Explicit feedback from Users
    - Implicit feedback from associated with user data
- Associating all the users and products for generating recommendations are often not meaningful

Dataset

Data Set



MovieLens

GroupLens Research has collected and made available rating data sets from the MovieLens web site (http://movielens.org). The data sets were collected over various periods of time, depending on the size of the set. Before using these data sets, please review their README files for the usage licenses and other details.

Help our research lab: Please take a short survey about the MovieLens datasets

MovieLens 100K Dataset

Stable benchmark dataset. 100,000 ratings from 1000 users on 1700 movies. Released 4/1998.

- README.txt
- ml-100k.zip (size: 5 MB, checksum)
- Index of unzipped files

Permalink: http://grouplens.org/datasets/movielens/100k/

MovieLens 1M Dataset

Stable benchmark dataset. 1 million ratings from 6000 users on 4000 movies. Released 2/2003.

- README.txt
- ml-1m.zip (size: 6 MB, checksum)

Permalink: http://grouplens.org/datasets/movielens/1m/

MovieLens 10M Dataset

Stable benchmark dataset. 10 million ratings and 100,000 tag applications applied to 10,000 movies by 72,000 users. Released 1/2009.

- README.html
- ml-10m.zip (size: 63 MB, checksum)

Permalink: http://grouplens.org/datasets/movielens/10m/

- MovieLens Dataset
- 100,000 Ratings (1 - 5)
- Available in different granularities.
- Structure:
  - 943 Users on 1682 Movies
  - Each user has rated at least 20 movies
  - Simple demographics info of users.

BI Use Case

BI use case : Problems with Original Algorithms

- The raw algorithms run on the whole database of users and items.
- Two typical problems:
  - Exponentially huge size of User-Item matrix.
  - Doesn’t leverage the association, similarity and dissimilarity between users and items.

Approach

Clustering- The Concept

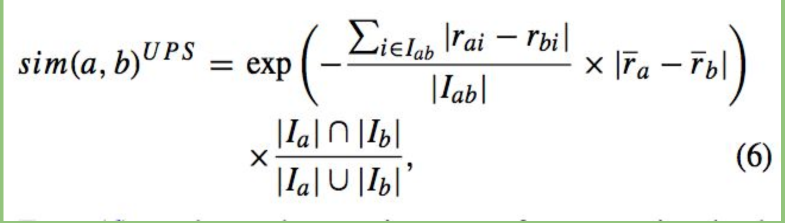
- Significance:
    - Speed
    - Size
    - Meaning of Recommendation

- Types:
    - User Clustering
    - Item Clustering

Optimist, Pessimist and Neutral Users

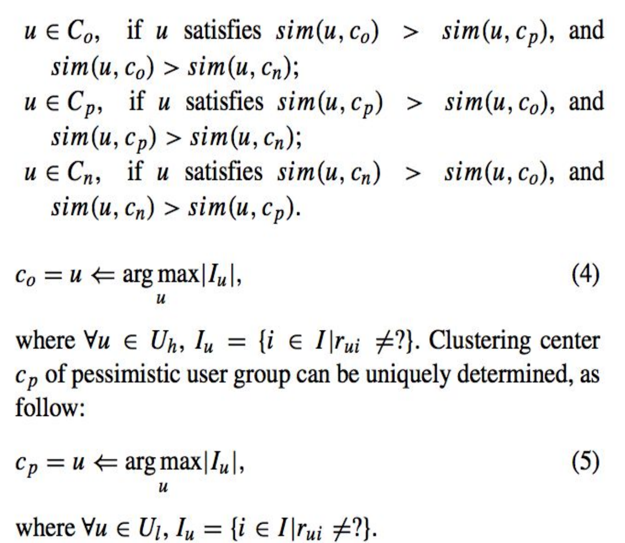
- Normally done using Demographics, but not good.
- Using user prior recommendation information.
- Set α and β, as defined in the paper.
- Select UL, UH and UN
- Find the leaders.

Similarity Matrices

- Baseline Matrices PCC and Cosine.
- New matrix UPS.
- 

$$sim(a, b)^{UPS} = \exp\left(-\frac{\sum_{i \in I_{ab}} |r_{ai} - r_{bi}|}{|I_{ab}|} \times |\bar{r}_a - \bar{r}_b|\right) \times \frac{|I_a| \cap |I_b|}{|I_a| \cup |I_b|}, \qquad (6)$$

Clustering



$u \in C_o$, if u satisfies $sim(u, c_o) > sim(u, c_p)$, and $sim(u, c_o) > sim(u, c_n)$;

$u \in C_p$, if u satisfies $sim(u, c_p) > sim(u, c_o)$, and $sim(u, c_p) > sim(u, c_n)$;

$u \in C_n$, if u satisfies $sim(u, c_n) > sim(u, c_o)$, and $sim(u, c_n) > sim(u, c_p)$.

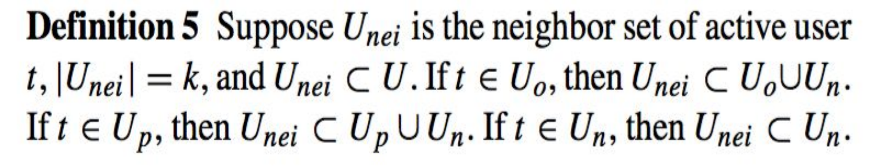
$$c_o = u \Leftarrow \arg\max_u |I_u|, \tag{4}$$

where $\forall u \in U_h$, $I_u = \{i \in I | r_{ui} \neq ?\}$. Clustering center c_p of pessimistic user group can be uniquely determined, as follow:

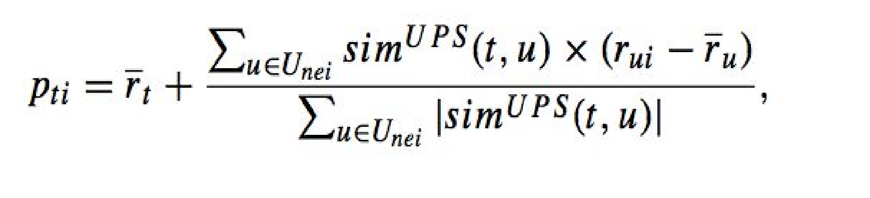
$$c_p = u \Leftarrow \arg\max_u |I_u|, \tag{5}$$

where $\forall u \in U_l$, $I_u = \{i \in I | r_{ui} \neq ?\}$.

Recommendations



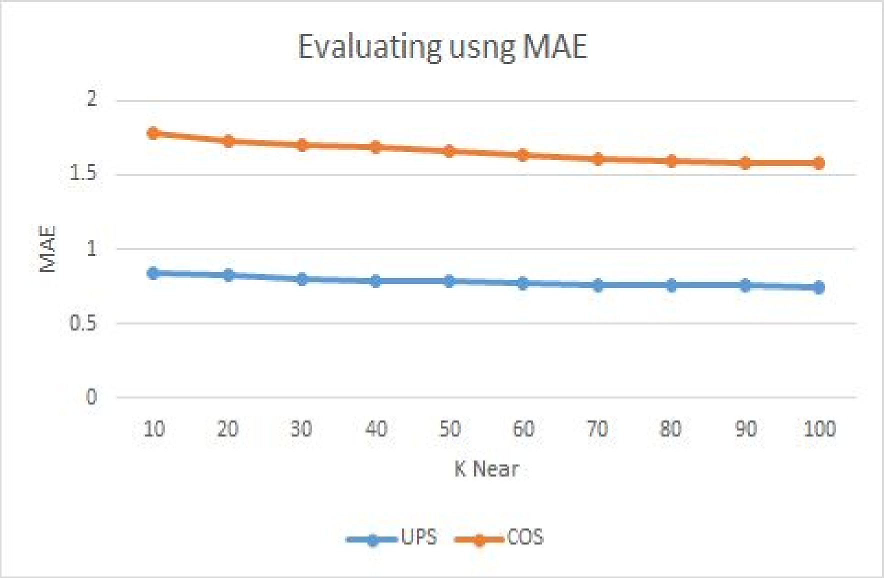
Definition 5 Suppose U_{nei} is the neighbor set of active user t, $|U_{nei}| = k$, and $U_{nei} \subset U$. If $t \in U_o$, then $U_{nei} \subset U_o \cup U_n$. If $t \in U_p$, then $U_{nei} \subset U_p \cup U_n$. If $t \in U_n$, then $U_{nei} \subset U_n$.



$$p_{ti} = \overline{r}_t + \frac{\sum_{u \in U_{nei}} sim^{UPS}(t, u) \times (r_{ui} - \overline{r}_u)}{\sum_{u \in U_{nei}} |sim^{UPS}(t, u)|},$$

Evaluation and Observation

Evaluation and observations



- $MAE = \frac{1}{\#U}\sum_{u\in U}\frac{\sum_{i\in I_u}|p_{ui} - r_{ui}|}{\#I_u}$
- UPS similarity metrics as compared to Cosine similarity metrics
- Optimal value of K

Conclusion and Future Work

Conclusion and Future Work

- Introducing the Bias into the recommendation system.
- How will it behave if done for items?
- Refine the criteria of optimism and pessimism.
- The people who give bad ratings on very famous and enjoyed items are the pessimists!