Admin

More project marking at CO339

- Today 12-12:30
- Tomorrow 9:30-10

Thursday presentations:

- Chuan Law, Jack Grunfeld, RuoHao Sun
- Three more speakers please!!!
- Very hard to find other times if you miss it

Project report due this Friday

- 1 page for marking
- 2 pages report using ChatGPT or other tools

Today: More on recommender systems

Typical Collaborative Filtering

- Memory based collaborative filtering
  - Nearest-neighbor based
    - User similarity
    - Item similarity
- Classification/Clustering for collaborative filtering
  - Model based
    - Naïve Bayes
    - Neural networks,
    - Kmeans
    - LDA (Latent Dirichlet Allocation, topic modeling)
    - LSA (Latent Semantic Analysis), Singular Value Decomposition (SVD)
  - Group oriented, less personalized, can be addressed by reducing cluster size

Content based filtering

- Content
  - Features:
    - Movie: directors, actor/actress, producers., editors, distributors, editors, keywords, review, ….
    - Text recommendation: a set of extracted keywords
  - Domain dependent
  - Getting features can be challenging

Classification/clustering problem

- like/dislike, similar/not_similar, user-specific classification
- Many algorithms can apply
  - NN-based, Deep learning as the state-of-art

Search problem,

- relevant/not relevant
- Focus on one user, Personalized search

Content based recommendation

- Tutorial:
  - Recommender Systems in Python: Beginner Tutorial
  - https://www.datacamp.com/community/tutorials/recommender-systems-python
- Movie dataset with metadata

- Tf-Idf representation
- Cosine similarity
- Find similar items

Recommender Systems Evaluation

- Consider ranking score

- MAE: mean absolute error

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|\,.$$

Hybrid

- Collaborative filtering:
  - Require other users rating data (cold start problem)
  - Can do cross domain
  - Non-transitive association problem: users are linked by common items and items are linked by common users.
- Content Based
  - Require one user's rating data
  - Require item's content data
  - Not cross domain
- Sequential/parallel Hybridization
- Combinational Hybridization

Limitations and extensions

- User similarities
  - Do not consider the relevance of items
  - Aware of item similarity
- User rating and prediction
  - Number between 1-5
  - Probability: (0.2, 0.3, 0.6, 0.1, 0.1)
  - Implicit: Watching time, thumb up/down, clicks, downloads, behavior
  - Explicit: User search history, user specified criteria
- Hybrid
  - Sequential, parallel
  - Diamond shape

In reality

- Many algorithms are used and then combined
- Netflix had a competition, and the winner used over 107 algorithms

- Many rankers, each has a different focus, each use many algorithms.

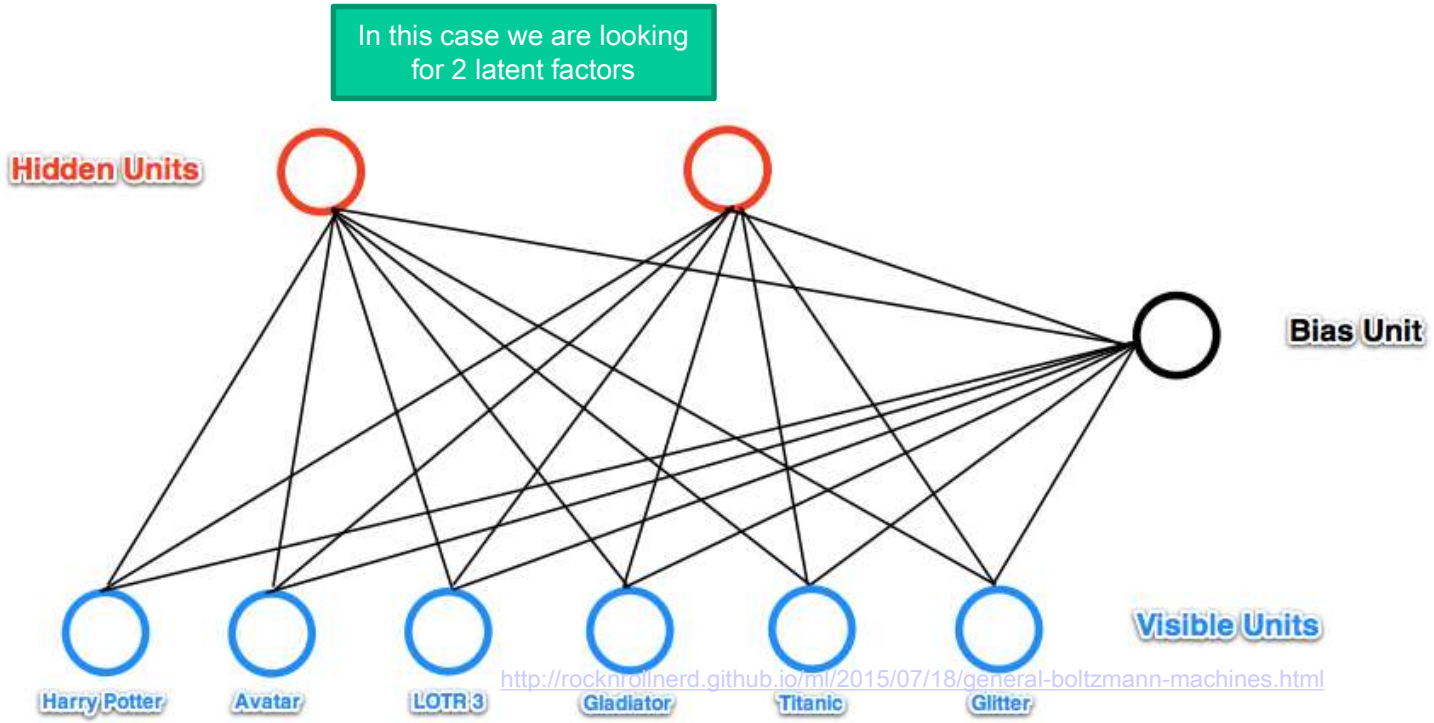
Algorithms Netflix uses

- Linear Regression
- Logistic Regression
- Elastic Nets
- Singular Value Decomposition
- Restricted Boltzmann Machines
- Markov Chains
- Latent Dirichlet Allocation
- Association Rules
- Gradient boosted decision trees
- Clustering such as K-Means
- Matrix factorization

One example:
Restricted Boltzmann Machines

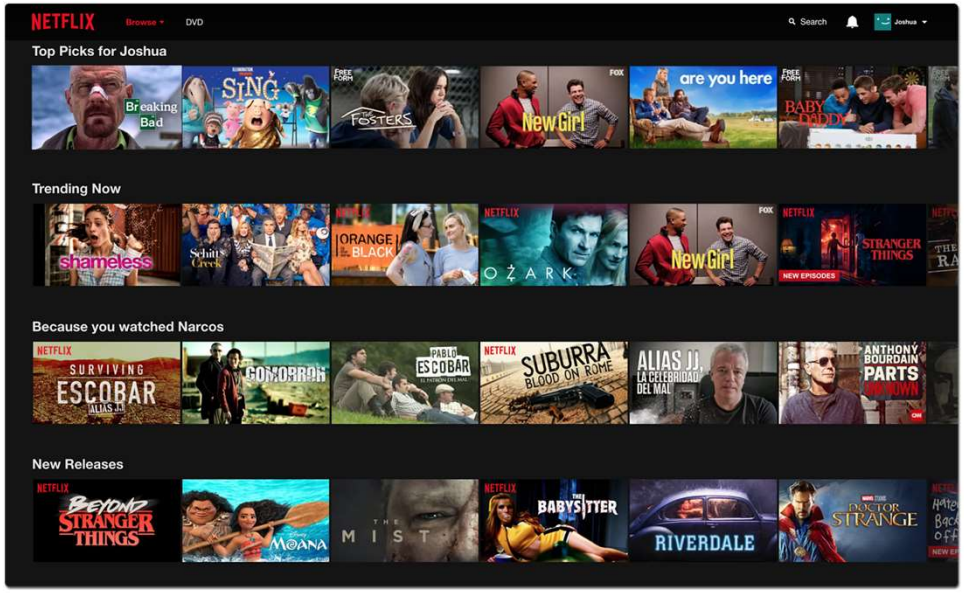
The learning algorithm will generate probabilities for if a user likes a piece of content or not, once it is trained (learned weights).

One layer of visible units, these are the the users' movie (or TV show) preferences which we know and set.
One layer of hidden units, the hidden connections we try to find ie the Pixar factor or maybe the film is award winning.
bias unit to account for movies' popularity.



Other Recommendations

- Netflix has many aspects of what combine to become its “Recommender System”
- Personalized Video Ranker PVR
- 40 rows on each home page
- Each may have up to 75 videos per row.
- Each row may have a different algorithm.
- For example under a given genre, different profiles will have those videos presented in a different order. PVR only gets to rank a standard list in a genre, the order is personalized but not the content

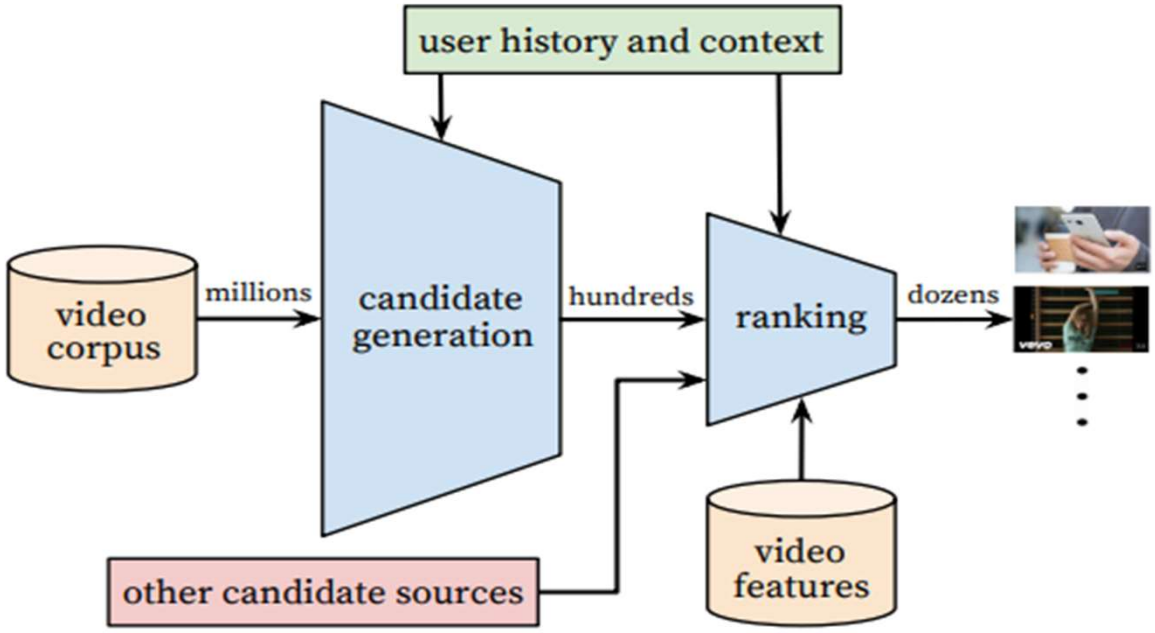


What does YouTube use?

- Association rules: finding relations between variables
  - Videos co-watched within 24 hours: linked videos
- User personal activity
  - Long watched videos
- Combine the two
  - Long watch videos link to more videos
  - Recursively link to more videos
  - Global video network
- Ranking:

  Video quality, user specificity, diversity

Personalized Recommendation



Association Rule Mining

YouTube's simplified algorithm used to score the relatedness of videos given a seed video, or in other words map a video v_i to a set of similar videos R_i.

"Association Rule Mining or Co-visitation counts"

$$r(v_i, v_j) = \frac{c_{ij}}{f(v_i, v_j)}$$

where c_i, c_j are the total occurrence counts for videos v_i and v_j
c_{ij} is the co-visitation count.
$f(v_i,v_j)$ is a normalization function (usually set as cj).

Ranking Measurements:

1) Video Quality: used to judge the likelihood that the video will be appreciated irrespective of the user (number of views, rating, comments, etc)
2) User Specificity
3) Diversification:Don't want all videos from the same channel or uploader

Using a linear combination of these three, we generate a personalized recommendation up to N videos to present to the user (Top-N Recommender).

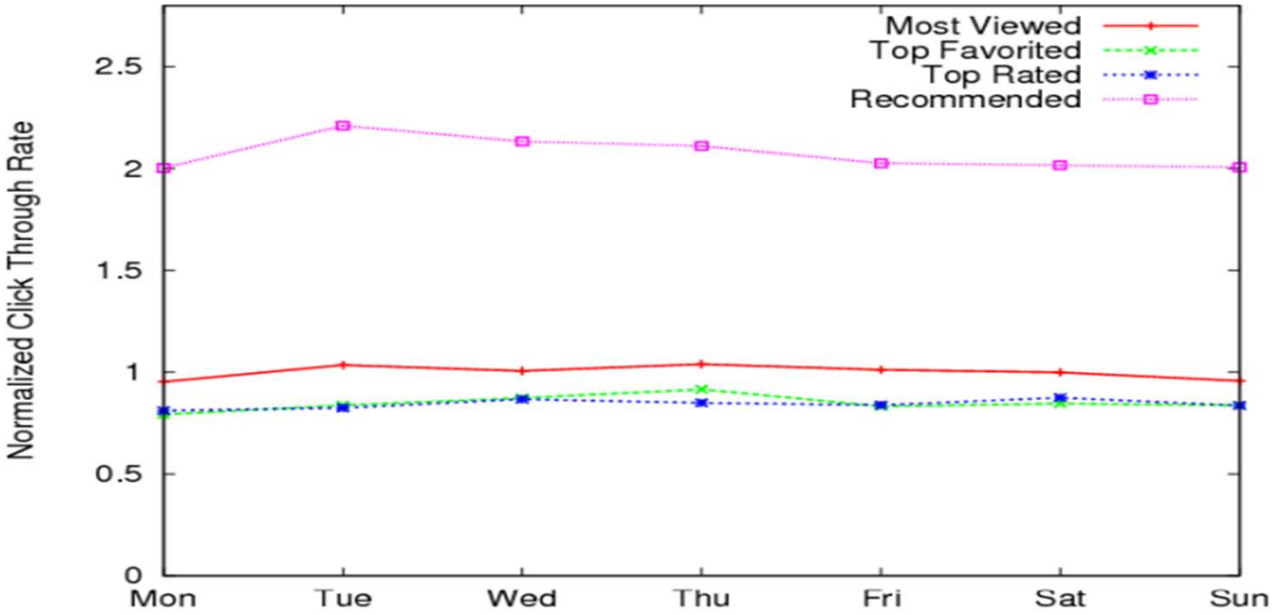
Evaluating Recommendation Quality

- CTR (Click Through Rates): is the ratio of the number of clicks on a video to the number of times that video was seen
- Long CTR: only counting clicks that led to watches of a substantial fraction of the video
- Session Length
- Time until first long watch
- Recommendation Coverage: The fraction of logged in users with recommendation.

The CTR for recommended videos exceeded Most Views, Top Rated, etc

Evaluation

Per-day average CTR for different browse page types over a period of 3 weeks



Evaluation in reality, in practise

- A/B testing
- A/B testing (sometimes called split testing) is experimenting and comparing two types or variations of an online or offline campaign such as a landing page, ad text, a headline, call-to-action or just about any other element of a marketing campaign or ad.

- By displaying two variations of your campaign, you can see which one attracts more interaction and conversions from your customers.
  - e.g. CTR (clickthrough rate): the number of clicks that your ad receives divided by the number of times your ad is shown