Collaborative Filtering
Machine Learning for Finance (FIN 570)

Instructor: Jaehyuk Choi

Peking University HSBC Business School, Shenzhen, China

2023-24 Module 3 (Spring 2024)

Recommendation Engine

- A recommendation system (recommender platform/engine) is a subclass of information filtering system that predicts the "rating" or "preference" a user would give to an item.
- Areas of usage
  - Online stores: recommend next items to buy.
  - Search engine: order search results for users in a personalized way.

A problem setup

- Set of users (buyers/subscribers)
- Set of items (movies, books, etc)
- Feedback information
  - Explicit: ratings, grades, etc.
  - Implicit: purchase, click, etc.
- Predict the score for missing user-item pairs.

$\boldsymbol{R}$	Starwars	Titanic	007
Choi	4	5	2
Oh	5	??	??
Sohn	??	3	4
Park	??	??	??

Contents-based vs collaborative filtering

Contents-based filtering

- Recommendations based on item descriptions/features and the preference/history of the “target” user.
- Equivalent to a user-specific classification problem (like vs hate).
- Suitable when item features (variables) and the target user’s history (training set) are rich.
- TF-IDF often used to extract the feature vectors.
- Example: Apps recommending rock music or romantic comedy movies.

Collaborative filtering (CF)

- Recommendations based on a user’s past behavior as well as the decisions made by similar users.
- Based on “people who agreed in the past will agree in the future.”
- The features of items are not required.
- Divided into memory-based vs model-based.

CF: memory-based approach

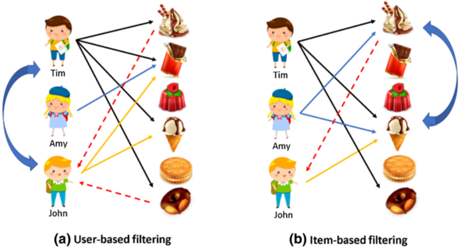
User-based

- Find user(s) who share the same rating patterns with the target user.
- Predict the rating on the item as the average ratings by the similar users.

Item-based

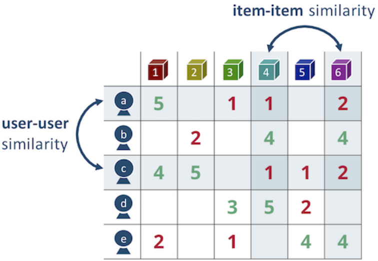
- Find product(s) with similar rating patterns by other users
- Predict the rating by the target user as the average ratings on the similar products.

- K-NN or kernel (cosine similarity) are calculated.
- The average can be replaced by the sum weighted by the similarity.



(a) User-based filtering

(b) Item-based filtering


item-item similarity
user-user similarity

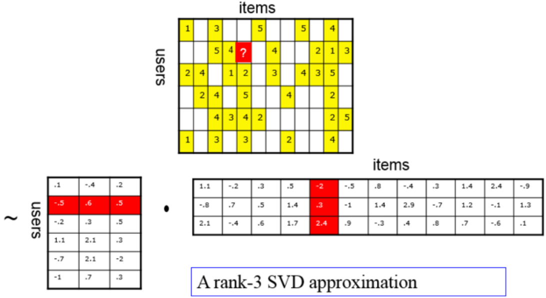
	1	2	3	4	5	6
a	5		1	1		2
b		2		4		4
c	4	5		1	1	2
d			3	5	2	
e	2		1		4	4

CF: matrix factorization (MF)

- Decomposes the user-item interaction matrix $\boldsymbol{R}$ into the product of two lower dimensionality rectangular matrices, $\boldsymbol{P}$ and $\boldsymbol{Q}$.

$$\boldsymbol{R} = \boldsymbol{P}\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{S}_k\boldsymbol{V}^T = \boldsymbol{U}\sqrt{\boldsymbol{S}_k} \cdot \sqrt{\boldsymbol{S}_k}\boldsymbol{V}^T.$$

- $\boldsymbol{P}$ is the sensitivity matrix of the (user, latent factor) pair.
- $\boldsymbol{Q}$ is the feature matrix of the (latent factor, item) pair.
- $\boldsymbol{P}$ and $\boldsymbol{Q}$ are solved to minimize the error, $\|\boldsymbol{R} - \boldsymbol{P}\boldsymbol{Q}\|_F$ with missing values.
- The algorithm is similar to the truncated SVD.



1		3			5			5		4	
		5	4	?		4			2	1	3
2	4		1	2		3		4	3	5	
	2	4		5			4			2	
		4	3	4	2					2	5
1		3		3			2			4	

.1	-.4	.2
-.5	.6	.5
-.2	.3	.5
1.1	2.1	.3
-.7	2.1	-2
-1	.7	.3

1.1	-.2	.3	.5	-2	-.5	.8	-.4	.3	1.4	2.4	-.9
-.8	.7	.5	1.4	.3	-1	1.4	2.9	-.7	1.2	-.1	1.3
2.1	-.4	.6	1.7	2.4	.9	-.3	.4	.8	.7	-.6	.1

A rank-3 SVD approximation

- The missing rating is predicted as

$$R_{ij} = P_{i*}Q_{*j} = \sum_k P_{ik}Q_{kj}.$$

- MF is a typical model-based CF.
- The ratings are explained by a few latent factors. For movies, it could be *action*, *romance*, *science fiction*, *classic vs contemporary*, *female vs male* etc. But we have to guess the latent factors.
- Depending on the model and loss function, MF is divided into **Funk MF**, **SVD++**, and **Asymmetric SVD**. See Wiki for detail.
- The algorithms are available in python.
- MC became widely known by Simon Funk during the Netflix prize in 2006 due to its effectiveness.
- Considered as the best single-model approach to CF.

- For summary, see
  Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. Computer, 42(8), 30–37. https://doi.org/10.1109/MC.2009.263

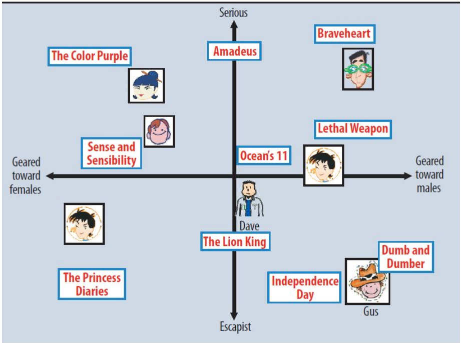


Figure 2. A simplified illustration of the latent factor approach, which characterizes both users and movies using two axes—male versus female and serious versus escapist.

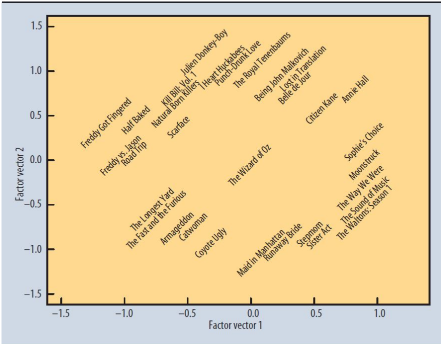


Figure 3. The first two vectors from a matrix decomposition of the Netflix Prize data. Selected movies are placed at the appropriate spot based on their factor vectors in two dimensions. The plot reveals distinct genres, including clusters of movies with strong female leads, fraternity humor, and quirky independent films.

Netflix Prize

- An open competition (2006-2009) for the best collaborative filtering algorithm to predict user ratings for movies, based on previous ratings without any other information about the users or movies (i.e., identified only by number IDs).
- Netflix provided a training data set of 100M ratings (user, movie, date, grade) that 480K users gave to 18K movies.
- The qualifying data set: 2.8M triplets (user, movie, date), with grades known only to the jury. Split to the quiz set and test set.
- The performance of algorithms measured by RMSE (5-points grade).

Netflix Prize Result

- In 2009, the grand prize of US$1M was given to the BellKor's Pragmatic Chaos, the combined team of *BellKor*, *Pragmatic Theory* and *BigChaos*. They achieved 10.09% improvement in RMSE.
  - Töscher, A., Jahrer, M., & Bell, R. M. (2009). The BigChaos solution to the netflix grand prize (pp. 1–52) [Netflix prize documentation]. Download.
  - Netflix did not implemented the method in the system due to model complexity. "We evaluated some of the new methods offline but the additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment."
- The Ensemble, the combined team of *Grand Prize Team*, *Opera Solutions and Vandelay United* ranked as 2nd place. They also achieved 10.09% improvement, but they submit the algorithm 20 minutes later.
- Simon Funk (matrix factorization) was ranked 10th place with 6.31% improvement.
- In 2010, Netflix canceled the 2nd competition over the privacy concern. "On December 17, 2009, four Netflix users filed a class action lawsuit against Netflix, alleging that Netflix had violated U.S. fair trade laws and the Video Privacy Protection Act. On March 19, 2010, Netflix reached a settlement with the plaintiffs, after which they voluntarily dismissed the lawsuit."

Matching problem in econ/finance

In finance/econ, there are many "matching" problems with "rating".

- M&A is the matching problem between the acquiring and acquired firm.
- Board directors and firms:
  - Erel, I., Stern, L. H., Tan, C., & Weisbach, M. S. (2021). Selecting Directors Using Machine Learning. The Review of Financial Studies, 34(7), 3226–3264. https://doi.org/10.1093/rfs/hhab050
  - They used machine learning methods with features combined from users and items.
  - Can we improve the performance by introducing MF?
- Banks and firms in corporate lending.
- Firm characteristics and macroeconomic market conditions to predict stock returns?