Recommender Systems Challenge 2017

Tommaso Scarlatti
897651

Overview

- **Application domain:** music streaming service, where users listen to tracks and create playlists
- **Goal:** discover which track a user will likely add to a playlist
- **Evaluation:** MAP@5 (Mean Average Precision)

$$AP@5 = \sum_{k=1}^{5} \frac{P(k) \cdot \mathrm{rel}(k)}{\min(m, 5)}$$

$$MAP@5 = \frac{\sum_{u=1}^{N} AP@5_u}{N}$$

- **57.561 | 10.000** playlists | target
- **100.000 | 32.195** tracks | target
- **1.040.522** interactions

Data Preprocessing

Pandas

- Read/write .csv
- Manage datasets efficiently
- Build up a validation set

Scikit-learn

The module *sklearn.preprocessing* was used to **binarize** the input data and then **normalize** the matrices

train_final.csv

```
playlist_id track_id
3271849 2801526
5616275 727878
11267488      2805283
10103900      1515105
3836898 2945623
5270369 2821391
3794808 1166185
7908370 2498280
11460733      282687
886396  863177
5758965 676462
```

Grouped by playlist

```
playlist_id
7569          [2634448, 2693660, 222710
7614          [2285204, 1384962, 371143
7641          [3825235, 257968, 3711129
7692          [3036454, 1439032, 302730
7816          [2305305, 1039409, 267481
7912          [126424, 218197, 3767380,
```

CSR matrix

```
(45646, 50383)      1
(45646, 87184)      1
(45646, 94907)      1
(45647, 15126)      1
(45648, 9984) 1
(45648, 46177)      1
(45648, 38592)      1
(45648, 1632) 1
(45648, 69508)      1
(45648, 30378)      1
```

Global strategies

Indices

- known_indices ✅
- non_target_indices ✅
- owner_indices ❌

Recommendations

- One playlist per cycle to avoid computation of large dense matrices

KNN

- K-nearest-neighbours used in every similarity matrix

Matrices

- Sparse csr matrix to speed up the dot product

Attributes

Playlists

- Only *owner_id* considered with no success
- Playlists of **URM** used as attributes to compute a similarity matrix

Tracks

- **Used:** *artist_id*, *album*, *tags*
- **Unused:** *duration*, *playcount*
- **77.040** total used attributes

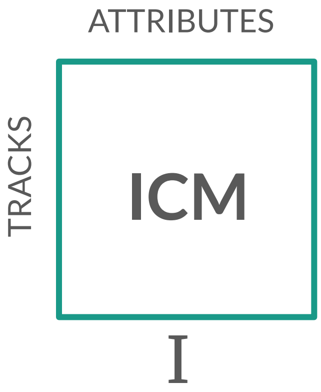
A first approach: Top-N recommender

- First naive attempt: a **non-personalized** recommender system
- Count each distinct track occurrence in *train_final.csv*
- Select the top 5 popular tracks
- Recommend these 5 tracks for all the target playlist

	playlist_id	track_ids
0	10024884	1563309 1363985 3705881 1595978 3166665
1	10624787	1563309 1363985 3705881 1595978 3166665
2	4891851	1563309 1363985 3705881 1595978 3166665
3	4267369	1563309 1363985 3705881 1595978 3166665
4	65078	1563309 1363985 3705881 1595978 3166665
5	10637124	1563309 1363985 3705881 1595978 3166665
6	3223162	1563309 1363985 3705881 1595978 3166665

MAP@5 = 0.001

Content-based recommender



- Item similarity matrix (cosine similarity)

$$S = II^T$$

- Recommendation: top 5 for similarity

$$\tilde{R} = RS$$

- **MAP@5**

0.01122 NO TF-IDF	0.05524 WITH TF-IDF	0.07695 TF-IDF + L2-NORM

Item-based collaborative filtering

- User content matrix (build from URM)

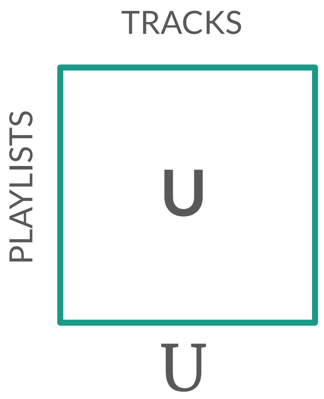
$$U = tfidf(R^T)^T$$

- Similarity matrix (cosine similarity)

$$S = U^T U$$

- **MAP@5** = 0.06653

$$\tilde{R} = RS$$



CB + CF recommender

- Added **l2 normalization** everywhere
- Weighted sum of S_ICM and S_U
- Much relevant S_ICM
- $\alpha \simeq 0.65$
- **MAP@5** = 0.09205

Combining predictions

α $\boxed{R_u \ \mathrm{ICM}}$

+

$(1-\alpha)$ $\boxed{R_u \ \mathrm{U}}$

SVD - Singular Value Decomposition

- New similarity matrix with ***k*** = 1000

  latent factors and **knn** = 250

- Computationally expensive ⟶ *scipy.sparse.linalg.svds*

- Very little improvements combining it with other recommenders

- **MAP@5** = 0.04553

$$\boxed{ICM} = \boxed{U}\ \boxed{S}\ \boxed{V^T}$$

Slim BPR - Bayesian Personalized Ranking

- Mainly based on the code on that we have seen in class
- **lil_matrix** to incrementally build the similarity matrix
- Added positive and negative item **regularization** terms
- Added **knn = 500**
- **Positive interactions** = number of non-zeros
- **MAP@5** = 0.04954

- learning rate = 0.01
- epochs = 1
- positive_reg = 1.0
- negative_reg = 1.0

Round robin & Ranking average

- Combines recommendation of: content-based, item-based collaborative filtering and SLIM
- Pick tracks according to their ranking

Round robin

- Tested in 3 different modes:
  "Standard", "Jump", "Mono"
- **MAP@5**: no improvements

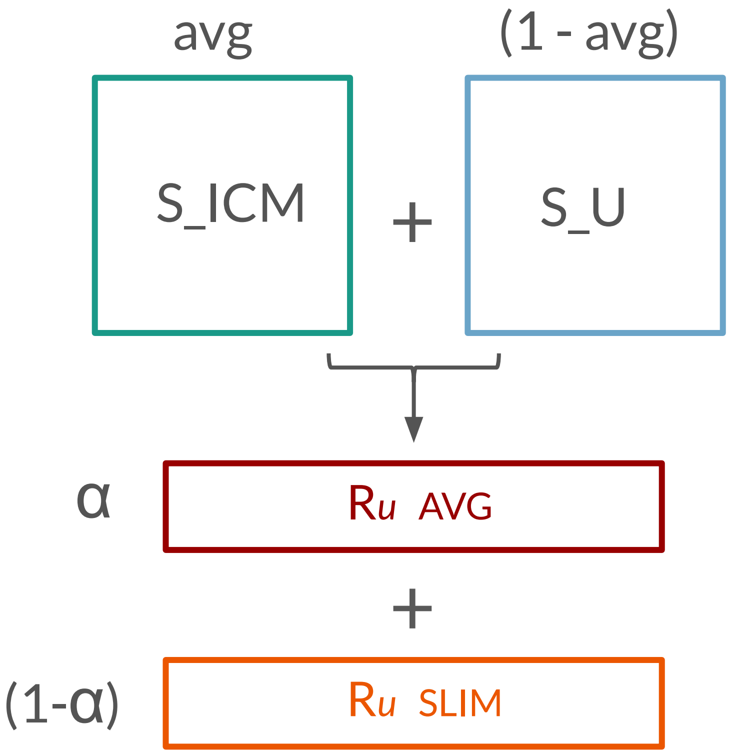
Ranking average

- Compute the ranking average for each track and pick the top 5 according to this value

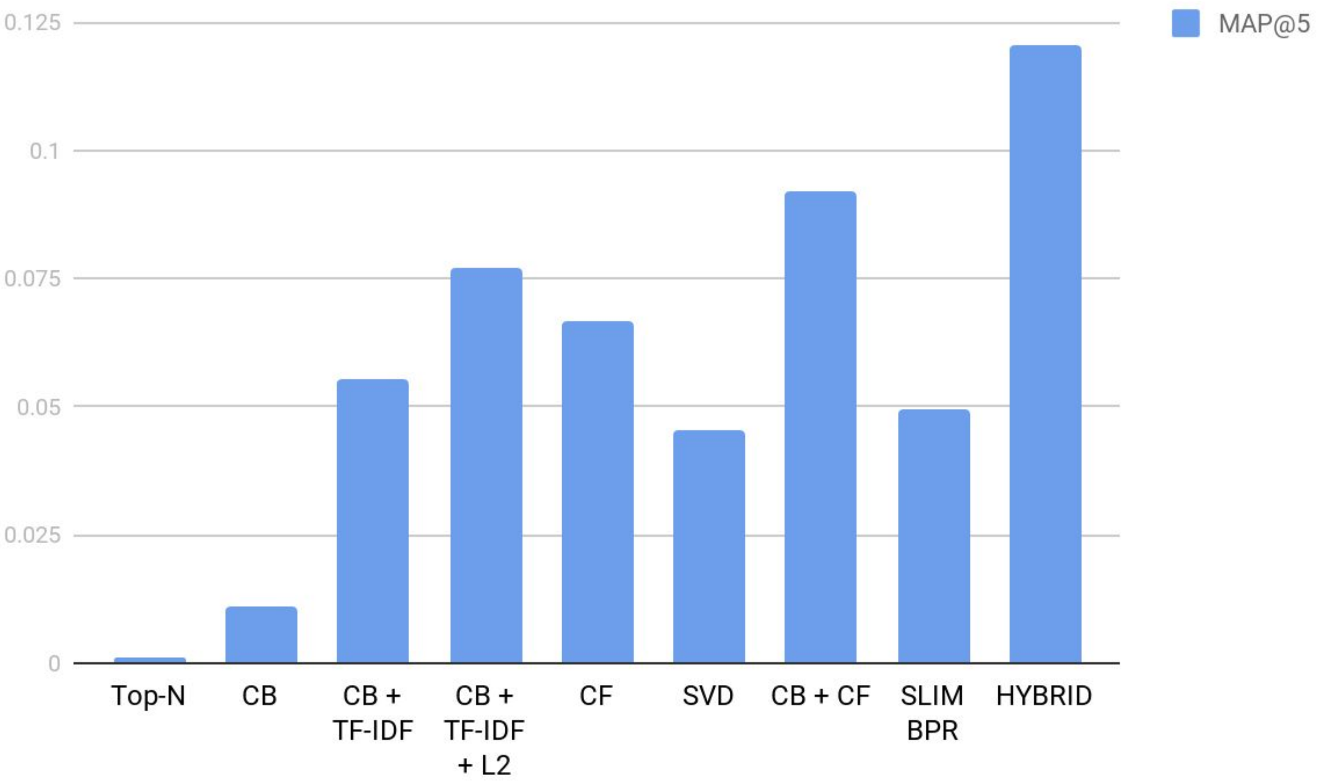
Hybrid Recommender (best solution)

- Merging models + combining predictions
- Apply tf-idf on the transpose of URM
- Merge similarity matrices
- Combine prediction with Slim BPR
- avg = 0.74   α = 0.20

- **MAP@5** = 0.10205



Summary



MAP@5
0.125
0.1
0.075
0.05
0.025
0
Top-N
CB
CB + TF-IDF
CB + TF-IDF + L2
CF
SVD
CB + CF
SLIM BPR
HYBRID

Testing

- Training set: 80 %
- Test set: 20%
- Playlist with at least 10 tracks for the test set

- Hyperparameters tuning → iterative search with descending granularity

Recommender System Challenge 2017

POLITECNICO
MILANO 1863

Thank you for your attention.

Tommaso Scarlatti