Collaborative filtering algorithms are prone to mainstream-taste bias

Pantelis P. Analytis
University of Southern Denmark

Philipp Hager
University of Amsterdam

Why is collaborative filtering not working for all?

- It is well documented that the performance of collaborative filtering varies among users [4].
- However, we currently have a limited understanding of performance variation between users.

Contributions:

- We systematically investigate how five staple collaborative filtering algorithms perform for individual users.
- We investigate which user attributes are predictive of the observed performance variation.

User-level evaluation

- **Methods:** KNN User-User / Item-Item, NMF, FunkSVD, EASE
- **Datasets:** MovieLens-1M, Faces, Jester
- **Protocol:** 5-fold nested cross-validation (stratified per user)
- **Metrics:**
  - **Performance:** nDCG, RMSE, fraction of concordant pairs (FCP)
  - **Variation:** Gini coefficient, diff between bottom 1% and top 1%

Performance variation between users

Dataset	Model	FCP				nDCG			
		mean	p1	diff	Gini	mean	p1	diff	Gini
MovieLens 1M	KNN User-User	0.6457	0.4327	0.3843	0.0574	0.9364	0.5501	0.4407	0.0202
	KNN Item-Item	0.6034	0.4224	0.3950	0.0651	0.9209	0.6062	0.3850	0.0225
	NMF	0.5999	0.4315	0.3961	0.0582	0.9240	0.6219	0.3696	0.0223
	FunkSVD	0.6496	0.4345	0.3794	0.0555	0.9393	0.5369	0.4539	0.0189
	EASE	0.6135	0.4302	0.3924	0.0543	0.9302	0.6074	0.3836	0.0206
Faces	KNN User-User	0.7096	0.4605	0.3406	0.0479	0.9014	0.5794	0.4113	0.0402
	KNN Item-Item	0.7046	0.4573	0.3435	0.0483	0.8958	0.5406	0.4542	0.0426
	NMF	0.6920	0.4144	0.3841	0.0499	0.8862	0.5457	0.4489	0.0477
	FunkSVD	0.7069	0.4491	0.3529	0.0489	0.8979	0.5551	0.4384	0.0433
	EASE	0.7056	0.4261	0.3736	0.0480	0.8987	0.5507	0.4436	0.0418
Jester	KNN User-User	0.6609	0.4501	0.3483	0.0491	0.9224	0.6948	0.2950	0.0251
	KNN Item-Item	0.6554	0.4467	0.3500	0.0499	0.9212	0.6947	0.2950	0.0256
	NMF	0.6121	0.4338	0.3572	0.0588	0.9026	0.6796	0.3103	0.0320
	FunkSVD	0.6527	0.4472	0.3507	0.0505	0.9204	0.6853	0.3042	0.0253
	EASE	0.6499	0.4420	0.3544	0.0501	0.9200	0.6909	0.2991	0.0243

All methods show substantial performance variation for different users across datasets and measures.

- The absolute difference in FCP between the top 1% and bottom 1% of users consistently exceeds 34% across models and datasets.
- All methods make recommendations to a non-negligible proportion of users that are worse-than-chance (FCP < 0.5).

Predicting algorithm performance for individual users

We train linear models to predict recsys performance for specific users using features from earlier work on performance prediction [3, 1] and decision science [2]:

- Mean user rating, rating variance, log rating count, log item popularity, user Gini, and mean item Gini.
- **Mean taste similarity:** Mean Pearson correlation of user ratings with all other users (i.e., how mainstream is the user's taste).
- **Taste dispersion:** Std. of the user's Pearson correlation with all other users (i.e., how consistent is the user's preference).

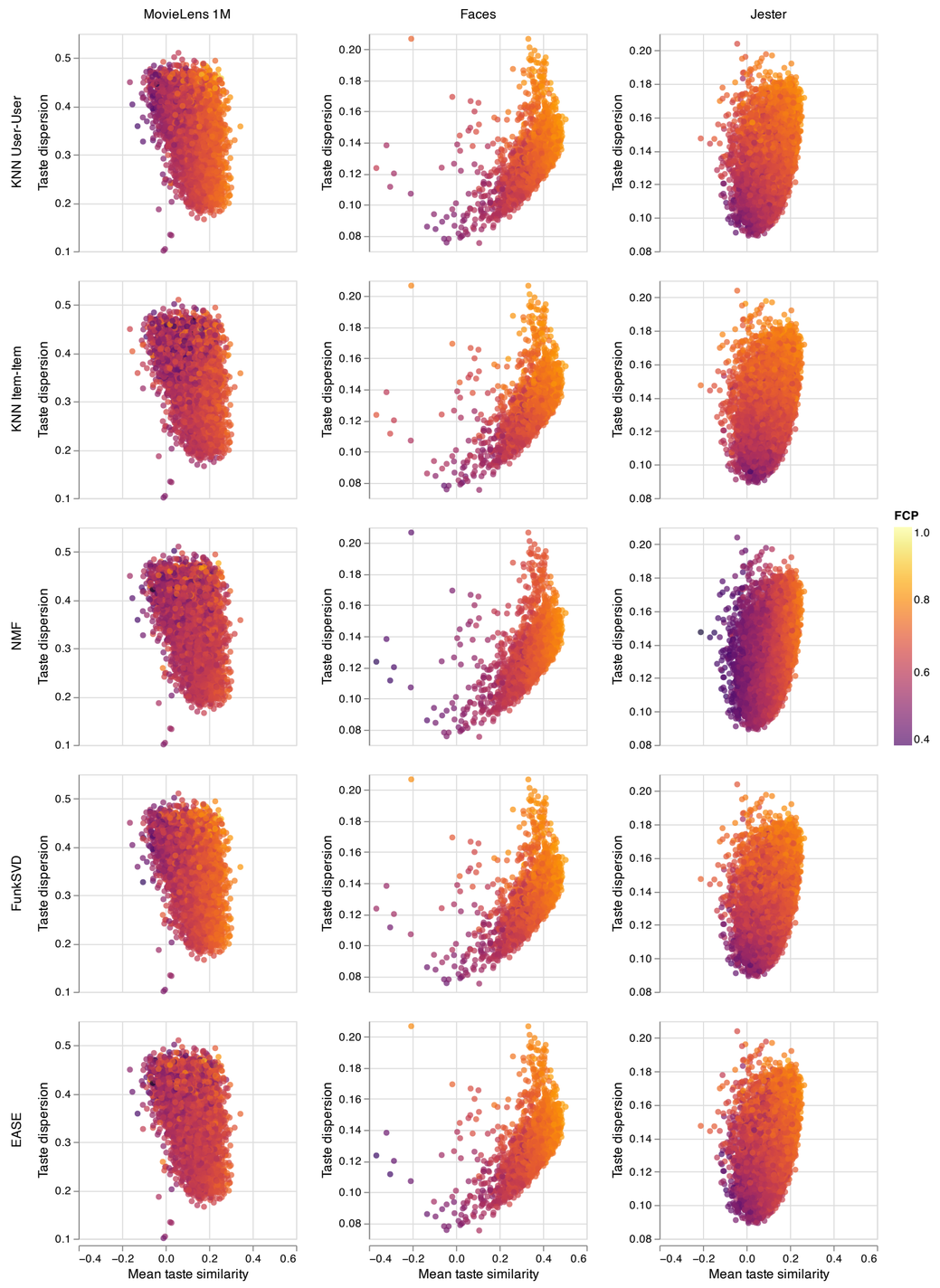
Dataset	Features	Model				
		User-User	Item-Item	NMF	FunkSVD	EASE
MovieLens	mean taste similarity	0.5204	0.2039	0.1214	0.4843	0.1345
	taste dispersion	0.0154	0.0545	0.0137	0.0213	0.0147
	mean taste similarity, taste dispersion	0.5443	0.2079	0.1212	0.4999	0.1344
	mean rating, rating variance, log rating count, log item popularity, user Gini, mean item Gini	0.0835	0.1816	0.03	0.118	0.043
	all	0.5601	0.2659	0.1349	0.5202	0.1505
Faces	mean taste similarity	0.6864	0.5665	0.8588	0.6593	0.7428
	taste dispersion	0.4536	0.5332	0.2039	0.4711	0.4317
	mean taste similarity, taste dispersion	0.7882	0.7227	0.8576	0.7541	0.8015
	mean rating, rating variance, log rating count, log item popularity, user Gini, mean item Gini	0.2274	0.2249	0.1665	0.2371	0.199
	all	0.8786	0.8326	0.9098	0.8591	0.8755
Jester	mean taste similarity	0.3699	0.2227	0.8623	0.3103	0.4372
	taste dispersion	0.5652	0.6768	0.118	0.5799	0.4499
	mean taste similarity, taste dispersion	0.6887	0.7077	0.8622	0.6654	0.6443
	mean rating, rating variance, log rating count, log item popularity, user Gini, mean item Gini	0.0015	0.0024	-0.0031	0.0146	0.0374
	all	0.7057	0.7213	0.8681	0.691	0.7135

Mean taste similarity and mean taste dispersion predict algorithm performance better than previously identified user features.

User taste predicts recommendation performance

We visualize algorithm performance using mean taste similarity and taste dispersion:

- Collaborative filtering excels for users with a mainstream taste.
- Users with alternative tastes receive worse recommendations (even worse than chance).
- Taste dispersion adds further nuance to the concept of mainstreamness by helping to identify specific groups of users (e.g. grey sheep users).



Conclusion

- Collaborative filtering performs vastly different across users and this difference can be predicted using taste profiles.
- Mean taste similarity and taste dispersion can explain a substantial portion of the performance variance.
- These features produce a mapping that unifies multiple previously proposed user categories (e.g., mainstream, grey sheep, or power users).
- User-level performance evaluation is crucial.

References

[1] Adomavicius, G., and Zhang, J. Impact of data characteristics on recommender systems performance. In *TMIS 2012*.

[2] Analytis, P. P., Barkoczi, D., and Herzog, S. M. Social learning strategies for matters of taste. In *Nature Human Behaviour 2018*.

[3] Ekstrand, M., and Riedl, J. When recommenders fail: predicting recommender failure for algorithm selection and combination. In *RecSys 2012*.

[4] Li, Y., Chen, H., Fu, Z., Ge, Y., and Zhang, Y. User-oriented fairness in recommendation. In *WWW 2017*.

pantelis@sam.sdu.dk, p.k.hager@uva.nl