Attribute-based Evaluation for RS: Incorporating User and Item Attributes in Evaluation Metrics

Pablo Sánchez, Alejandro Bellogín

pablo.sanchezp@uam.es, alejandro.bellogin@uam.es

Information Retrieval Group, Department of Computer Science, Universidad Autónoma de Madrid

IR Group @UAM

Universidad Autónoma de Madrid

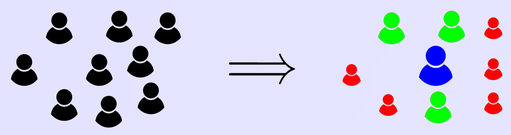
Motivation

- Most offline Recommender Systems evaluations focus on comparing how many recommended items are in the ground-truth. However, this evaluation methodology is **incomplete**
- **At user level**, we are ignoring the intrinsic characteristics of the users. Some of them may belong to less represented groups or may have less interactions than others (cold-start). **How these different user groups affect the recommenders performance?**
- **At item level**, there may be recommended items that although not found in the ground-truth, share a high similarity with the ones in the test set. **How can we take them into account?**

User and item level attribute recommendation

The common formulation for any recommendation metric can be expanded to incorporate a weight $c(u)$ for each user u:

$$m(r_u) = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} m(r_u,u) \Rightarrow m(r_u) = \frac{1}{\sum_u c(u)}\sum_{u\in\mathcal{U}} c(u)m(r_u,u) \quad (1)$$



We can set different weights for the recommended items that appear in the test set ($I^+(u)$), for the ones that have a non-zero similarity ($I^*(u)$) and the rest of the items ($I^-(u)$):

$$m(r_u,u) \propto \sum_{i\in I^+(u)} w^+(u,i) + \sum_{i\in I^*(u)} w^*(u,i) + \sum_{i\in I^-(u)} w^-(u,i) \quad (2)$$

Experiments and Results

User Attributes. NDCG@5 — Movielens

Rec	Std	Gender		Age				Training Quartile				Test Quartile			
		F	M	1	18	35	56	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Rnd	0.004	0.004	0.004	0.002	0.005	0.004	0.003	0.002	0.002	0.004	0.009	0.002	0.001	0.004	0.009
Pop	0.141	0.093	0.159	0.108	0.168	0.116	0.072	0.066	0.082	0.136	0.278	0.064	0.084	0.141	0.275
UB	0.300	0.250	0.320	0.258	0.326	0.279	0.254	0.210	0.243	0.310	0.444	0.205	0.248	0.316	0.444
IB	0.241	0.196	0.259	0.220	0.259	0.227	0.186	0.142	0.192	0.250	0.381	0.142	0.194	0.254	0.380
HKV	**0.315**	**0.266**	**0.335**	**0.276**	**0.336**	**0.300**	**0.261**	**0.218**	**0.260**	**0.330**	**0.465**	**0.213**	**0.262**	**0.336**	**0.465**
BPR	0.235	0.202	0.249	0.213	0.253	0.220	0.196	0.144	0.176	0.251	0.378	0.140	0.181	0.255	0.379
CB	0.018	0.014	0.020	0.013	0.019	0.017	0.020	0.010	0.014	0.022	0.027	0.009	0.015	0.023	0.026
CBCF	0.249	0.204	0.267	0.208	0.276	0.226	0.184	0.135	0.187	0.262	0.414	0.138	0.186	0.266	0.412
Sky	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Foursquare (New York City)

Rec	Std	Gender		Training Quartile				Test Quartile			
		F	M	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Rnd	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Pop	0.088	0.088	0.087	0.115	0.091	0.083	0.069	0.093	**0.099**	0.079	0.079
UB	**0.096**	**0.093**	**0.097**	0.111	**0.099**	**0.095**	**0.087**	**0.095**	0.097	**0.094**	**0.097**
IB	0.037	0.030	0.041	0.043	0.039	0.037	0.032	0.037	0.041	0.037	0.035
HKV	0.081	0.077	0.083	0.090	0.086	0.078	0.075	0.079	0.081	0.080	0.083
BPR	0.087	0.087	0.088	**0.117**	0.092	0.083	0.069	0.094	0.096	0.080	0.080
CB	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CBCF	0.074	0.074	0.074	0.076	0.088	0.074	0.069	0.072	0.078	0.069	0.079
Sky	0.997	0.997	0.996	0.997	0.997	0.995	0.997	1.000	0.992	0.992	0.998

User attributes

- Gender attribute: higher results in males (although there are more males than females)
- Age attribute: dominated by most frequent age groups (18 and 35)
- Cold start in training/test: different behaviour in the training/test quartiles in the two data sets
- In general: ranking of recommenders remains almost the same

Item Attributes NDCG@5

- $\tau = 0$: standard metric
- τ_m: main feature
- τ_s: secondary feature
- $\tau_{m,s}$: using main and secondary features

Movielens
τ_m = directors, τ_s = genres

Rec	$\tau = 0$	τ_m	τ_s	$\tau_{m,s}$
Rnd	0.004	0.013	0.191	0.196
Pop	0.141	0.164	0.277	0.288
UB	0.300	0.318	0.423	0.432
IB	0.241	0.260	0.370	0.380
HKV	**0.315**	**0.335**	**0.442**	**0.452**
BPR	0.235	0.255	0.374	0.384
CB	0.018	0.030	0.214	0.220
CBCF	0.249	0.267	0.391	0.400
Sky	1.000	1.000	1.000	1.000

Foursquare (NYC)
τ_m = level 3, τ_s = level 1 categories

Rec	$\tau = 0$	τ_m	τ_s	$\tau_{m,s}$
Rnd	0.000	0.031	0.221	0.230
Pop	0.088	0.098	0.205	0.208
UB	**0.096**	**0.126**	0.244	0.252
IB	0.037	0.081	**0.246**	**0.258**
HKV	0.081	0.110	0.225	0.233
BPR	0.087	0.098	0.204	0.206
CB	0.000	0.048	0.216	0.225
CBCF	0.074	0.100	0.218	0.224
Sky	0.997	0.997	0.997	0.997

Item attributes

- Higher results considering more features (decreased sparsity)
- Some recommenders change their relative performance: IB gets promoted when using the more generic (L1) categories in Foursquare, whereas Pop gets demoted
- Open problem: how should the weights for these extended test items be tuned?

Conclusions

- By exploiting user attributes we can detect if the data show specific biases towards different user groups or characteristics of the users [2]
- Using the item attributes we may be able to obtain a more complete vision about the recommenders. Similar conclusions were obtained in the user study in [1]
- Several issues are still open in RS evaluation, here we show how we can exploit user and item attributes to make an in-depth study about the recommendations produced

References

[1] Shir Frumerman, Guy Shani, Bracha Shapira, Oren Sar Shalom Are All Rejected Recommendations Equally Bad?: Towards Analysing Rejected Recommendations. In UMAP (2019), pp. 157–165.

[2] Alan Said and Alejandro Bellogín Coherence and inconsistencies in rating behavior: estimating the magic barrier of recommender systems. In User Model User-Adap Inter (2018) 28, pp. 97–125.

Source code available at: `https://bitbucket.org/PabloSanchezP/attreval4recsys`

RecSys 2019, 13th ACM Conference on Recommender Systems. September 16-20, 2019. Copenhagen, Denmark.

Work funded by the Ministerio de Ciencia, Innovación y Universidades (ref: TIN2016-80630-P) and by the European Social Fund, within the 2017 call for predoctoral contracts