Measuring anti-relevance: a study on when recommendation algorithms produce bad suggestions

Pablo Sánchez, Alejandro Bellogín

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

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

Motivation

- Most evaluation in RS area focus on measure the relevance of recommendations
- Bad recommendations can undermine the confidence that a user has in a system [1]
- How do we measure **bad** suggestions?

https://imgur.com/dEmDw

https://ericjohnbaker.files.wordpress.com/2015/01/wrecks-micky.png

Anti-Metrics: Adding Anti-Relevance to RS Evaluation

The **PRP** states that *if a system's response to a query is a ranking of documents in order of decreasing probability of relevance, the overall effectiveness of the system to its users will be maximized* [2].

$$m(R_u|\theta_{rel}) = C \sum_{i \in R_u} m(\theta_{rel}(r_{ui})|u,i)$$

We study the **dual PRP** problem (estimating the probability of anti-relevance and ranking the documents according to the opposite of this probability):

$$\overline{m}(R_u|\theta_{arel}) = C \sum_{i \in R_u} (1 - \overline{m}(\theta_{arel}(r_{ui})|u,i)) \propto 1 - C' \sum_{i \in R_u} m(\theta_{arel}(r_{ui})|u,i) = 1 - m(R_u|\theta_{arel})$$

Balance the results of relevance metrics and anti-relevance metrics with:

- The average $\mu(x) = \frac{1}{2}(x + \overline{x})$
- The harmonic mean $H(x) = 2\frac{x\overline{x}}{x+\overline{x}}$
- Taking the likelihood ratio $LH(x) = \frac{x}{1-\overline{x}}$

Experiments and Results

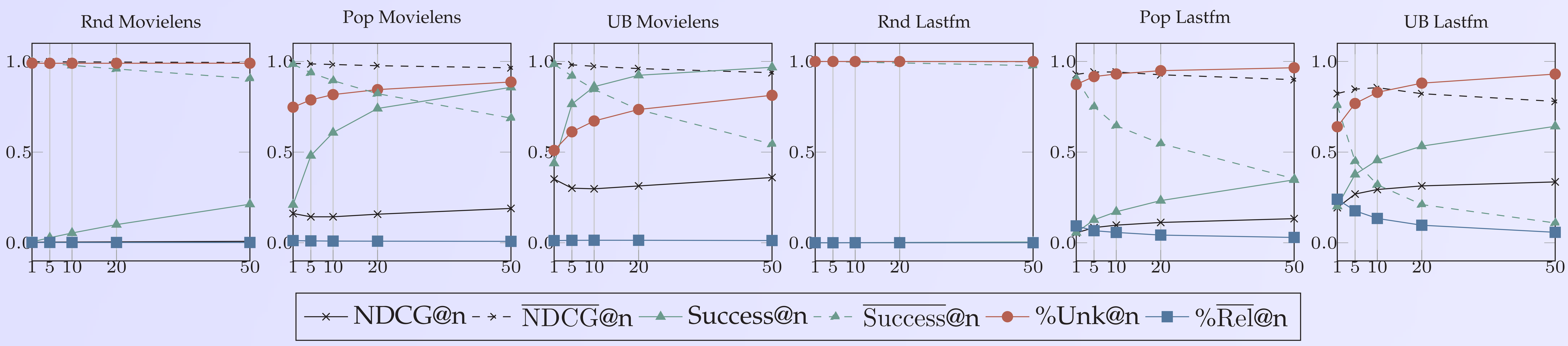
Comparing metrics and anti-metrics

Movielens: Results @10

Rec	NDCG	MAP	$\overline{\text{NDCG}}$	$\overline{\text{MAP}}$	%Rel	%$\overline{\text{Rel}}$	%Brd	%Unk
Rnd	0.004	0.002	‡0.998	‡0.999	0.6	‡0.2	†0.3	99.0
Pop	0.143	0.093	†0.983	†0.990	14.7	†0.9	2.7	81.7
IB	0.253	0.181	0.974	0.985	24.5	1.5	4.9	69.3
UB	†0.298	†0.211	0.973	0.984	†27.0	1.4	4.5	67.2
HKV	‡0.321	‡0.230	0.977	0.987	‡29.6	1.2	4.4	†65.0
BPRMF	0.246	0.178	0.962	0.978	24.9	2.2	5.5	67.5
Skyline	**1.000**	**1.000**	**1.000**	**1.000**	**79.2**	**0.0**	**0.0**	**0.0**
$\overline{\text{Skyline}}$	0.000	0.000	0.000	0.000	0.0	44.6	‡0.0	‡0.0

Lastfm: Results @10

Rec	NDCG	MAP	$\overline{\text{NDCG}}$	$\overline{\text{MAP}}$	%Rel	%$\overline{\text{Rel}}$	%Brd	%Unk
Rnd	0.000	0.000	‡1.000	‡1.000	0.0	‡0.0	†0.0	100.0
Pop	0.097	0.077	†0.942	†0.972	1.8	†5.7	0.5	93.1
IB	0.248	0.204	0.857	0.918	4.5	13.0	1.6	83.7
UB	†0.294	†0.248	0.855	0.917	†5.0	13.4	1.7	†83.0
HKV	‡0.316	‡0.272	0.875	0.931	‡5.2	11.9	1.7	84.4
BPRMF	0.240	0.195	0.875	0.932	4.4	11.8	1.5	85.1
Skyline	**0.984**	**0.980**	**1.000**	**1.000**	**12.8**	**0.0**	**0.0**	**0.0**
$\overline{\text{Skyline}}$	0.000	0.000	0.093	0.128	0.0	76.2	‡0.0	‡0.0



Case study using Lastfm: benchmarking similar methods. Results @5

Rec	Params	NDCG	MAP	$\overline{\text{NDCG}}$	$\overline{\text{MAP}}$	%Rel	%$\overline{\text{Rel}}$	%Brd	%Unk	μ(NDCG)	μ(MAP)	H(NDCG)	H(MAP)	LH(NDCG)	LH(MAP)
IB	(VC, 90)	**0.218**	**0.193**	0.856	0.884	0.025	0.161	0.021	**0.794**	0.537	0.538	0.347	**0.317**	1.515	1.662
HKV	(50, 0.1, 100)	0.215	0.188	**0.917**	**0.941**	0.024	**0.097**	**0.016**	0.862	**0.566**	**0.565**	**0.348**	0.314	**2.585**	**3.215**
BPRMF	(100, 1, 0.005)	0.213	0.183	0.870	0.902	**0.025**	0.149	0.021	0.805	0.542	0.542	0.342	0.304	1.640	1.860
UB	(SJ, 40)	**0.266**	**0.235**	0.846	0.877	**0.030**	0.176	0.025	**0.769**	0.556	0.556	**0.405**	**0.371**	1.732	1.907
HKV	(100, 0.1, 1)	0.263	0.232	**0.859**	**0.892**	0.030	**0.160**	**0.024**	0.786	**0.561**	**0.562**	0.403	0.368	**1.873**	**2.147**

Conclusions

- We have derived a framework to define anti-relevance metrics
- Unpersonalized recommenders tend to retrieve less anti-relevant items than personalized recommenders, although they also retrieve less relevant items since for these algorithms unknown items account for a large number of their recommendations
- HKV and UB are the best in terms of balance between anti-relevance metrics and classic relevance metrics
- What do we prefer, more uncertainty or anti-relevant items?

References

[1] JONATHAN L. HERLOCKER, JOSEPH A. KONSTAN, LOREN G. TERVEEN, JOHN RIEDL Evaluating collaborative filtering recommender systems. In ACM Trans. Inf. Syst. (2004), pp. 5–53.

[2] S. E. ROBERTSON The Probability Ranking Principle in IR. In Readings in Information Retrieval (1997), pp. 281–286.

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

RecSys 2018, 12th ACM Conference on Recommender Systems. October 2-7, 2018. Vancouver, Canada.