RecSys 2018, Vancouver, Canada

On the Robustness and Discriminative Power of IR Metrics for Top-N Recommendation

Daniel Valcarce* A. Bellogín† J. Parapar* P. Castells†

@dvalcarce @abellogin @jparapar @pcastells

*University of A Coruña

†Universidad Autónoma de Madrid

Evaluation

Recommender Systems Evaluation

Online evaluation (e.g., A/B testing):

- expensive,
- measures real user behavior.

Offline evaluation:

- cheap,
- highly reproducible,
- usually constitutes the first step before deploying a recommender system.

Recommender Systems Evaluation

Online evaluation (e.g., A/B testing):

- expensive,
- measures real user behavior.

Offline evaluation ←

- cheap,
- highly reproducible,
- usually constitutes the first step before deploying a recommender system.

Offline Evaluation of Recommender Systems (RS)

When evaluating recommender systems (RS), which **metric** should we use?

- Many types: error, ranking accuracy, diversity, novelty, etc.
- **Ranking accuracy** metrics are becoming the most popular.
- These metrics have been traditionally used in Information Retrieval (IR).

IR Metrics used in RS

Some IR metrics that have been used in RS:

- P: Precision
- Recall
- MAP: Mean Average Precision
- nDCG: Normalised Discounted Cumulative Gain
- MRR: Mean Reciprocal Rank
- bpref: Binary Preference
- infAP: Inferred Average Precision

Analyse how IR Metrics behave in RS

These ranking accuracy metrics have been studied in IR.

We want to study their behavior in top-N recommendation.

Two perspectives:

- robustness,
- discriminative power.

Proposal

Robustness

Sparsity bias

- Sparsity is intrinsic to the recommendation task.
- We take random subsamples from the test set to increase the bias.

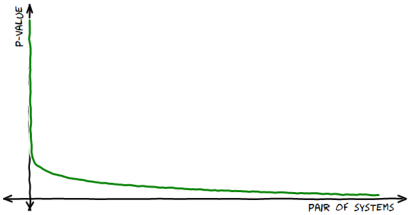
Popularity bias

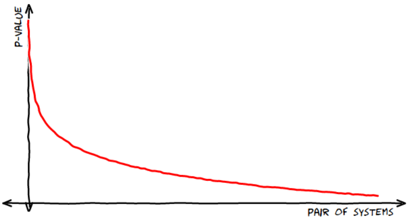
- Missing-not-at-random (long tail distribution).
- We remove the most popular items to study the bias.

We measure the robustness of a metric by computing the Kendall's correlation of systems rankings when changing the amount of bias.

Discriminative Power

- A metric is discriminative when its differences in value are statistically significant.
- We use the **permutation test** with difference in means as test statistic.
- We run a statistical test between all possible system pairs.
- We plot the obtained p-values sorted by decreasing value.





Experiments

Experimental Settings

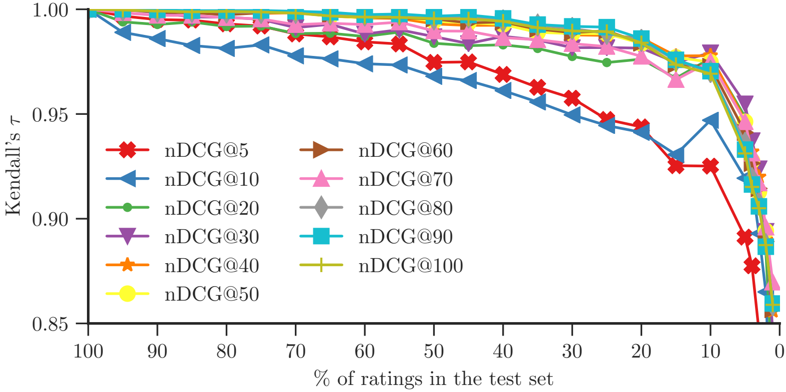
- Three datasets:
  - MovieLens 1M,
  - LibraryThing,
  - BeerAdvocate.
- Methodology:
  - AllItems: rank all the items in the dataset that have not been rated by the target user.
  - 80-20% random split.
- Systems:
  - 21 different recommendation algorithms.

Comparing metric cut-offs

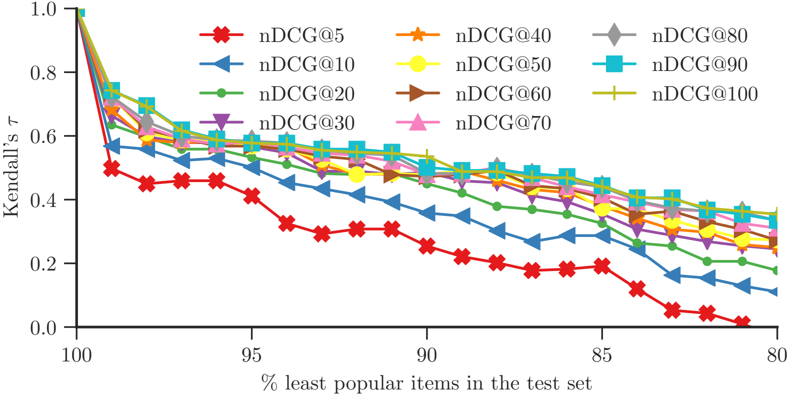
Comparing cut-offs of the same metric (nDCG) 1/4

	@5	@10	@20	@30	@40	@50	@60	@70	@80	@90	@100
@5	1.00	0.95	0.93	0.92	0.92	0.92	0.92	0.91	0.90	0.90	0.90
@10	0.95	1.00	0.98	0.97	0.97	0.97	0.97	0.96	0.95	0.95	0.95
@20	0.93	0.98	1.00	0.99	0.99	0.99	0.99	0.98	0.97	0.97	0.97
@30	0.92	0.97	0.99	1.00	1.00	1.00	1.00	0.99	0.98	0.98	0.98
@40	0.92	0.97	0.99	1.00	1.00	1.00	1.00	0.99	0.98	0.98	0.98
@50	0.92	0.97	0.99	1.00	1.00	1.00	1.00	0.99	0.98	0.98	0.98
@60	0.92	0.97	0.99	1.00	1.00	1.00	1.00	0.99	0.98	0.98	0.98
@70	0.91	0.96	0.98	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99
@80	0.90	0.95	0.97	0.98	0.98	0.98	0.98	0.99	1.00	1.00	1.00
@90	0.90	0.95	0.97	0.98	0.98	0.98	0.98	0.99	1.00	1.00	1.00
@100	0.90	0.95	0.97	0.98	0.98	0.98	0.98	0.99	1.00	1.00	1.00

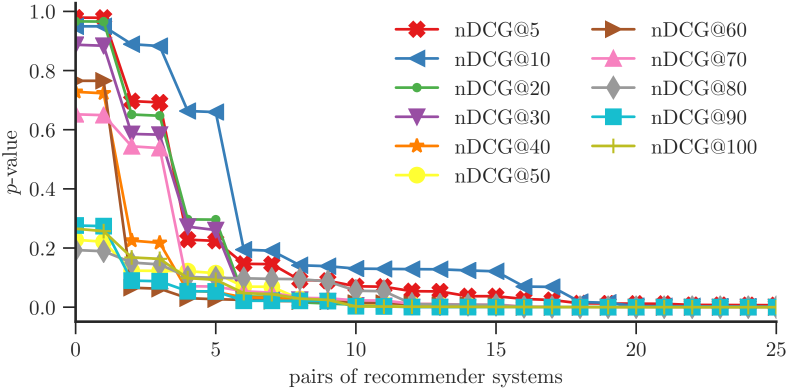
Correlation between cut-offs of nDCG.



Kendall's correlation among systems when increasing the **sparsity bias** using nDCG.



Kendall's correlation among systems when changing the **popularity bias** using nDCG.



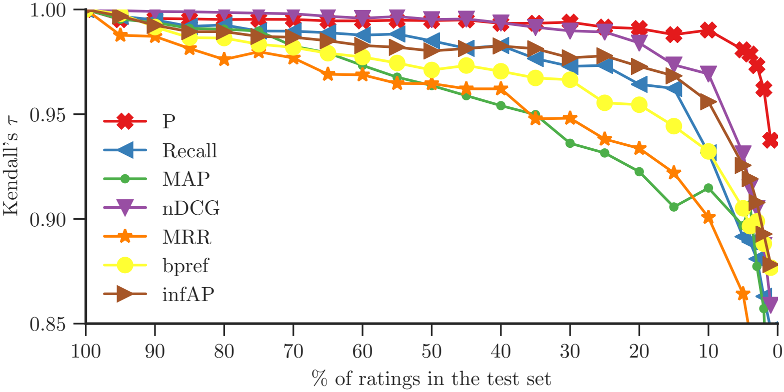
Discriminative power of nDCG measured with p-value curves.

Comparing metrics at the same cut-off

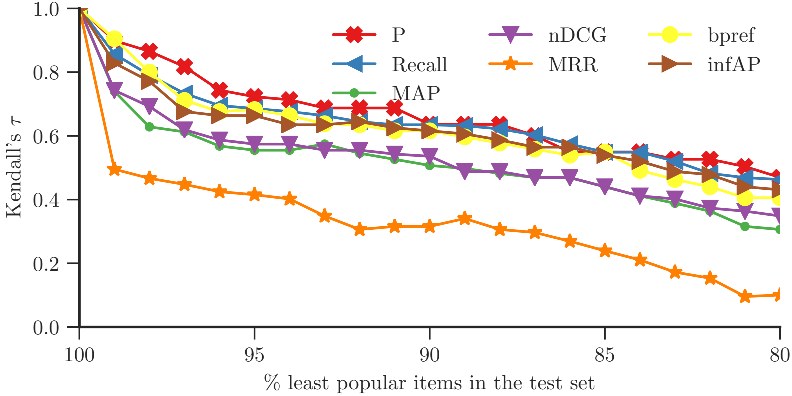
Comparing metrics with cut-off @100

	P	Recall	MAP	nDCG	MRR	bpref	infAP
P	1.00	0.89	0.87	0.89	0.71	0.89	0.91
Recall	0.89	1.00	0.87	0.90	0.72	0.90	0.92
MAP	0.87	0.87	1.00	0.96	0.84	0.92	0.92
nDCG	0.89	0.90	0.96	1.00	0.82	0.94	0.96
MRR	0.71	0.72	0.84	0.82	1.00	0.80	0.80
bpref	0.89	0.90	0.92	0.94	0.80	1.00	0.96
infAP	0.91	0.92	0.92	0.96	0.80	0.96	1.00

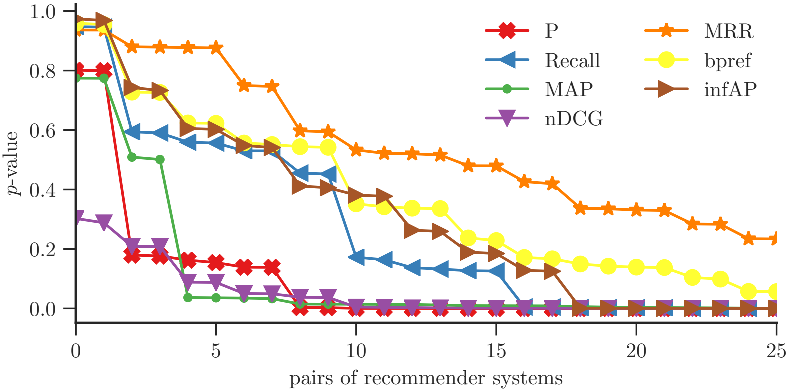
Correlation between metrics.



Kendall's correlation among systems when increasing the **sparsity bias**.



Kendall's correlation among systems when increasing the **popularity bias**.



Discriminative power measured with p-value curves.

Conclusions and Future Directions

Conclusions

- **Deeper cut-offs** offer **greater** robustness and discriminative power than shallow cut-offs.
- **Precision** offers **high** robustness to sparsity and popularity biases and **good** discriminative power.
- **NDCG** provides the **best** discriminative power and **high** robustness to the sparsity bias and **moderate** robustness to the popularity bias.

Future work

- Explore different types of metrics such as **diversity** or **novelty** metrics.
- Use **other evaluation methodologies** instead of AllItems.
    - For instance, One-Plus-Random: one relevant item and N non-relevant items as candidate set.
- Employ different partitioning schemes such as **temporal splits**.

Thank you!