Revealing the Hidden Impact of Top-N Metrics on Optimization in Recommender Systems

<u>Lukas Wegmeth</u>

Tobias Vente

Lennart Purucker

46th European Conference on Information Retrieval

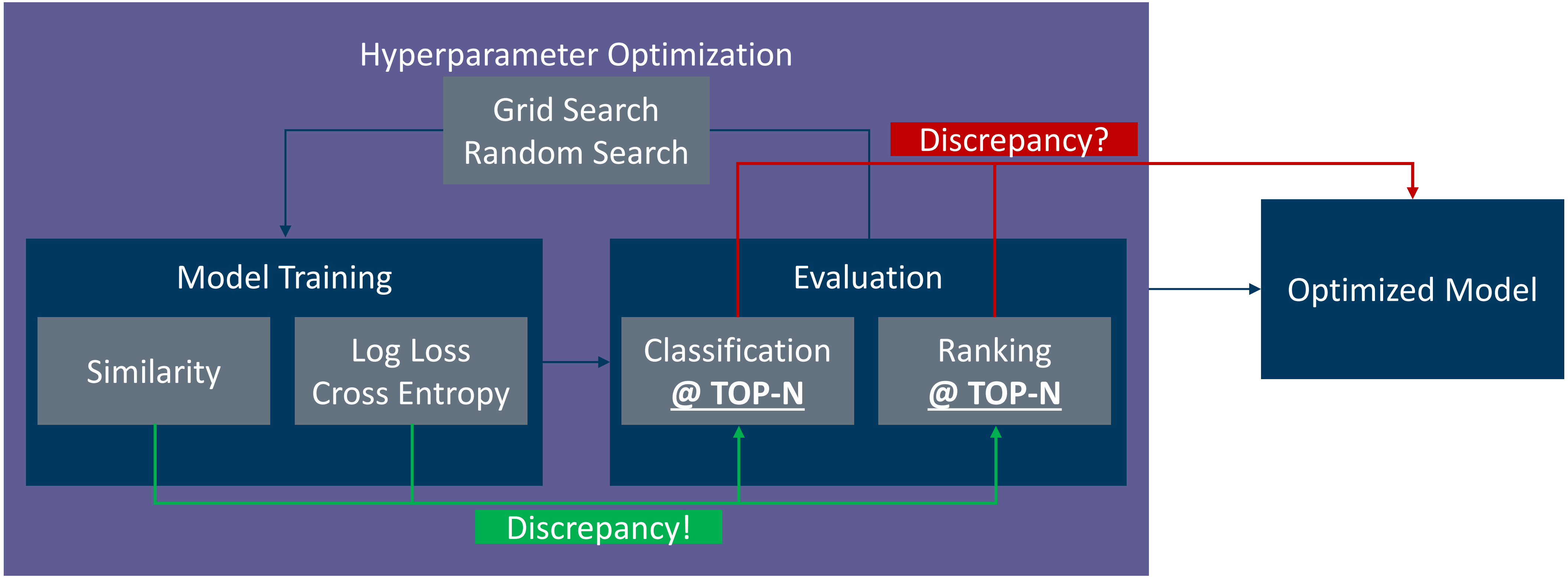
Lukas.Wegmeth@uni-siegen.de

@LWegmeth

www.isg.beel.org

www.uni-siegen.de

Problem


Hyperparameter Optimization
Grid Search
Random Search
Discrepancy?
Model Training
Similarity
Log Loss
Cross Entropy
Evaluation
Classification
@ TOP-N
Ranking
@ TOP-N
Optimized Model
Discrepancy!

Selection Strategies

P_{U_x}: *Ranked list that contains predicted items* I_y *for User* x

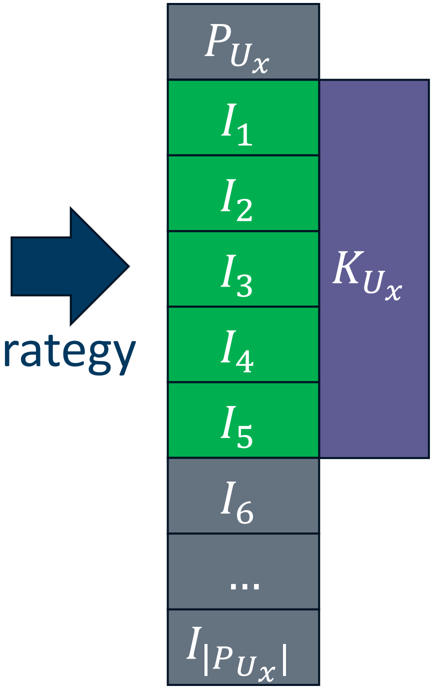
K_{U_x}: *Subset of* P_{U_x} *that contains the top* k *predicted items*

n: *Number of selected items for the evaluation of ranking metrics*

Typical selection to evaluate <ranking-metric>@5

$k = 5\,; \mathrm{n} = 5$

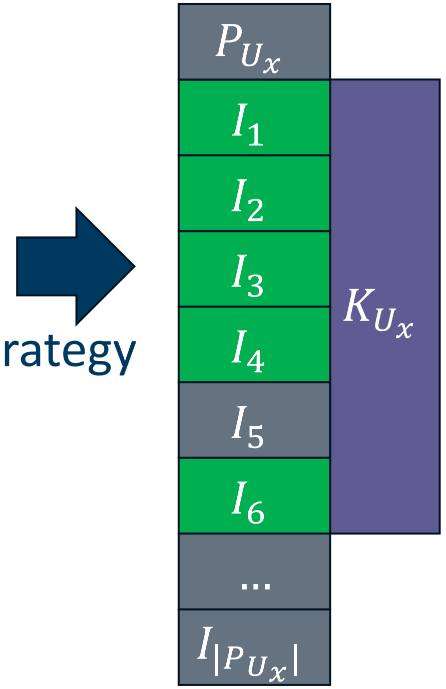
<u>Top-n</u>
selection strategy



Alternative selection to evaluate <ranking-metric>@5

$k = 6\,; \mathrm{n} = 5$

<u>Non-top-n</u>
selection strategy



$\binom{k}{n}$-1 Non-top-n selection strategies!

Research Questions

RQ1 Does the selection of items other than the top-n during the evaluation of recommender systems yield improved predictive accuracy for specific algorithms, domains, or data sets?

RQ2 If there are cases where selecting items other than the top-n improves predictive accuracy, is there a significant impact of top-n metrics on optimization?

Exploratory Study: Method

- We set $k = \mathbf{10}$ and $n = \mathbf{5}$ ➔ **251** non-top-n selection strategies + **1** top-n selection strategy

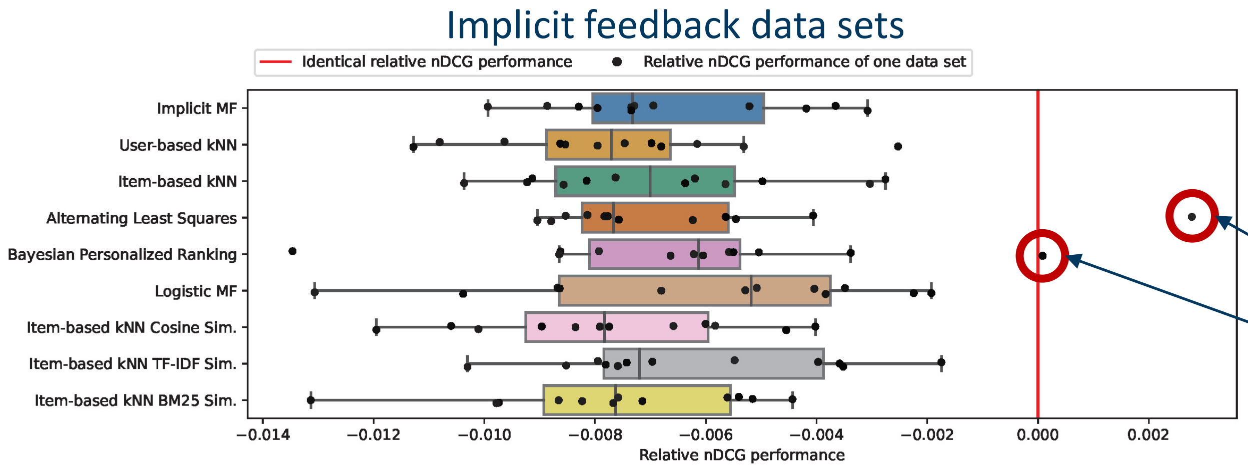
- **12** implicit feedback data sets
- **8** explicit feedback data sets

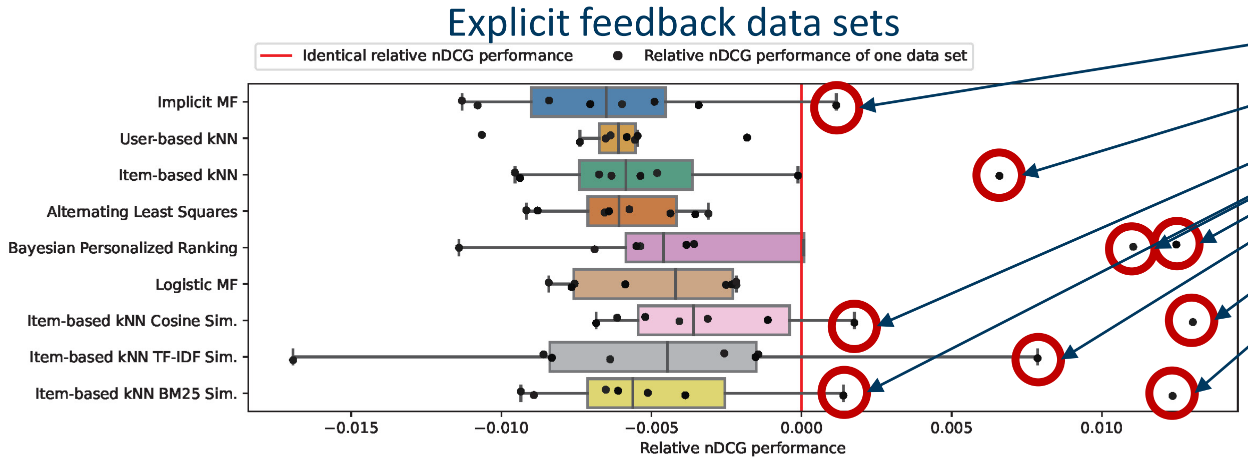
- **9** collaborative filtering recommendation algorithms (memory-based and model-based)
- **2** baseline recommendation algorithms (popularity and random)

- **5**-core pruning
- **60/20/20** user-based random split
- **5**-fold cross-validated

- Hyperparamter optimization with Random Search for **2** hours

Exploratory Study: Results Part 1



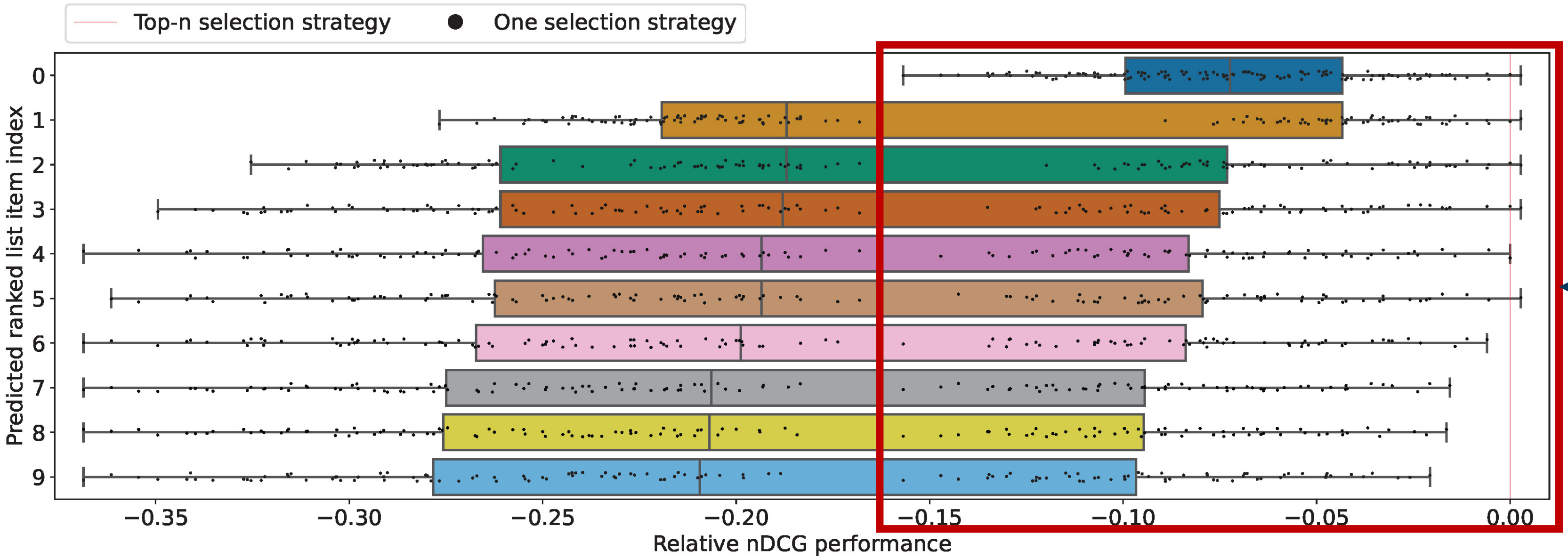


RQ1 Does the selection of items other than the top-n during the evaluation of recommender systems yield improved predictive accuracy for specific algorithms, domains, or data sets?

~0.01% of non-top-n selection strategies outperform the top-n selection strategy.

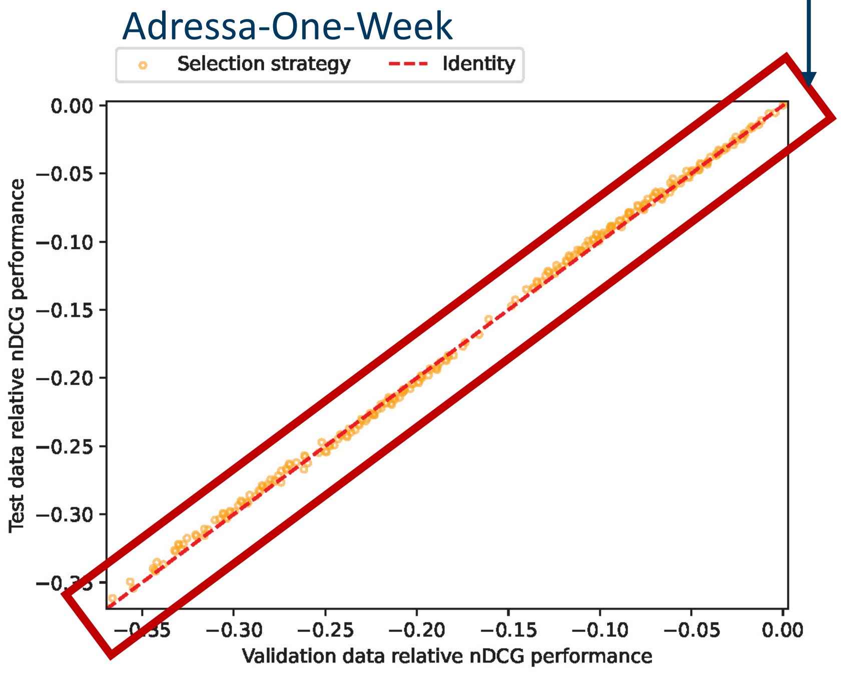
Exploratory Study: Results Part 2

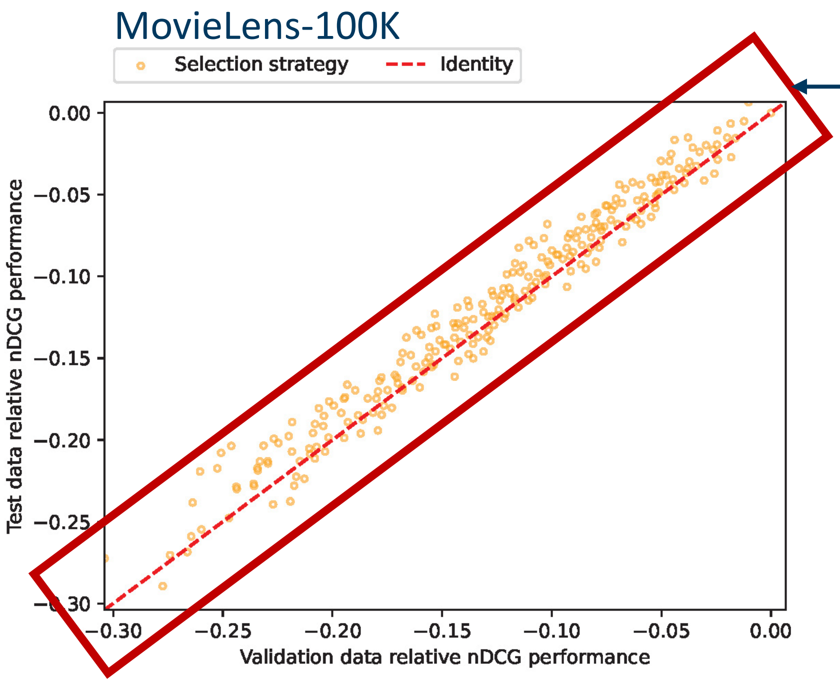
RQ2 If there are cases where selecting items other than the top-n improves predictive accuracy, is there a significant impact of top-n metrics on optimization?



~43% of the best selection strategies are not significantly different

Exploratory Study: Results Part 3

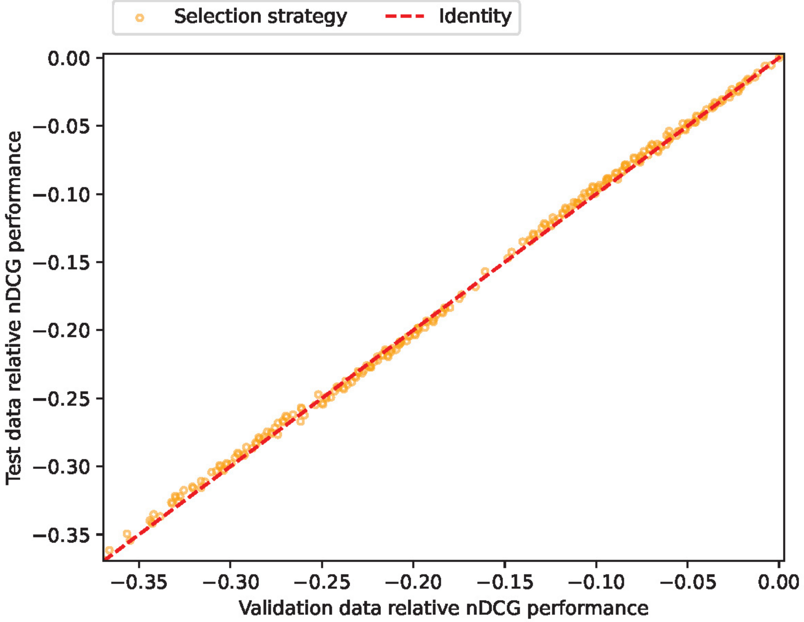




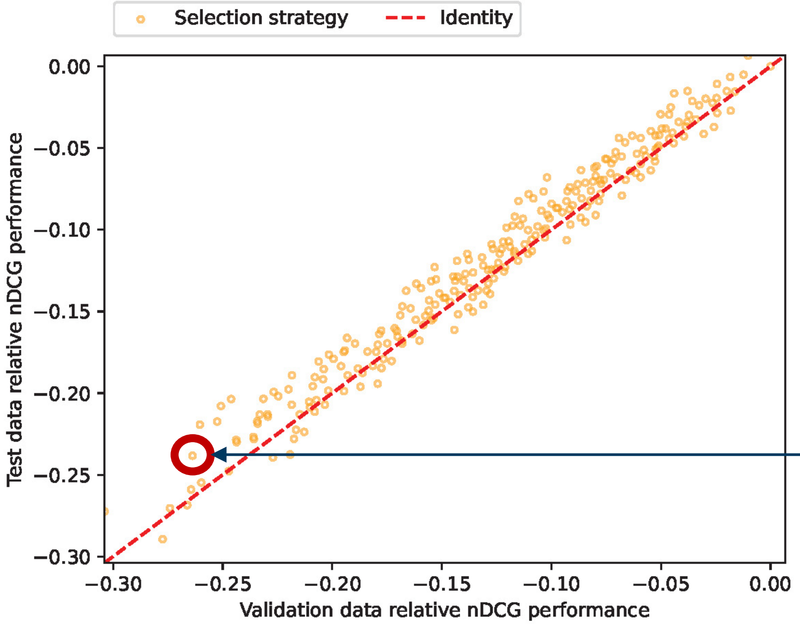
Selection strategies generalize from validation to test (Pearson > 0.99)

Exploratory Study: Results Part 4

Adressa-One-Week



MovieLens-100K



Expensive to calculate

➔ $\binom{k}{n}$ selection strategies

➔ $\binom{k}{n}$ x <ranking-metric>

Conclusion

We cleared any doubts of the hidden impact of top-n metrics on optimization as a confounding factor for the evaluation and reproducibility of traditional collaborative filtering algorithms.

We found no evidence indicating a practical benefit in optimizing with selection strategies other than the top-n!

Links

Revealing the Hidden Impact of Top-N Metrics on Optimization in Recommender Systems

https://code.isg.beel.org/scoring-optimizer

https://isg.beel.org/people/lukas-wegmeth/

https://isg.beel.org/people/tobias-vente/

https://ml.informatik.uni-freiburg.de/profile/purucker/