Sturgeon and the Cool Kids

Problems with Top-*N* Recommender Evaluation

Michael D. Ekstrand — *People and Information Research Team*

Boise State University

Vaibhav Mahant — *Texas State University*

https://goo.gl/bfVg1T

What can editorials in mid-20th-century sci-fi mags tell us about evaluating recommender systems?

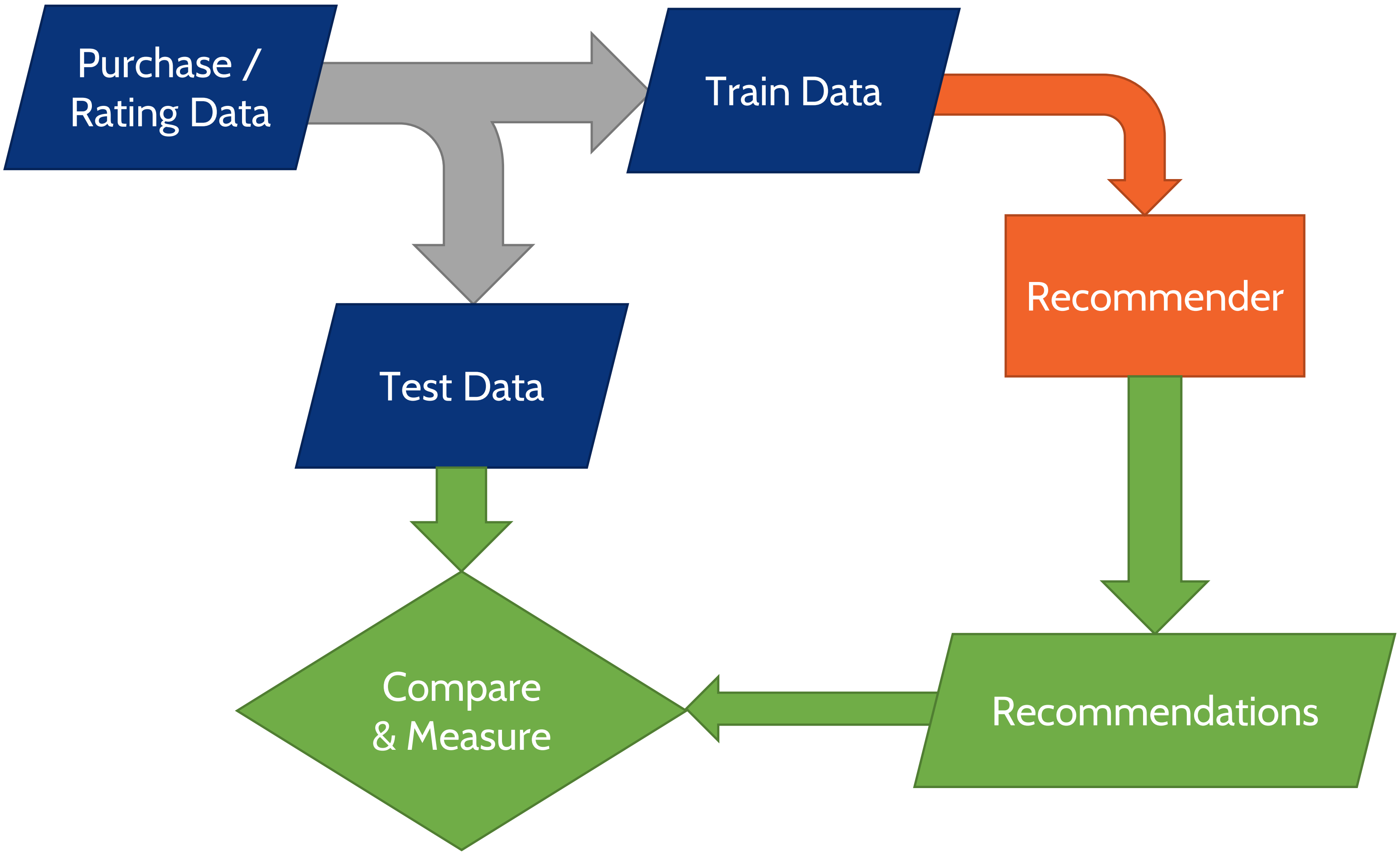
Evaluating Recommenders

Recommenders find **items** for **users**.

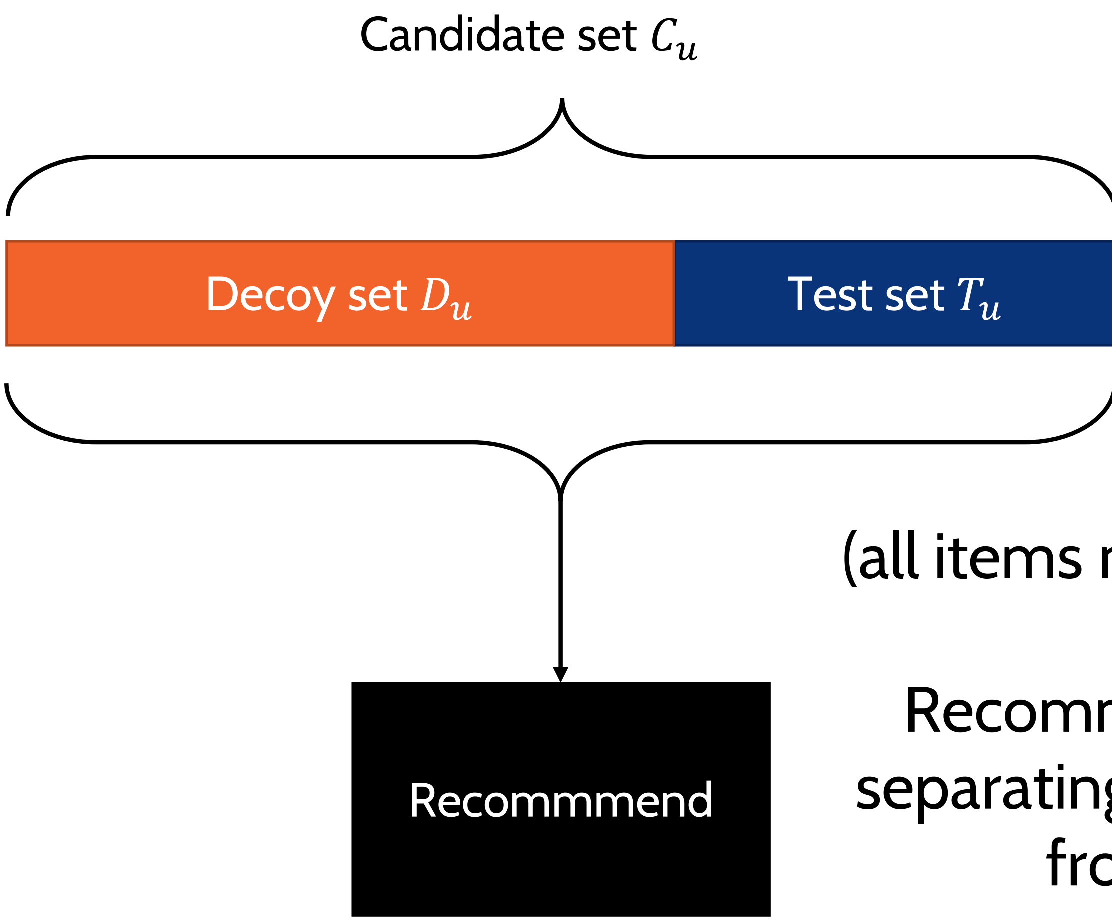
Evaluated:

- **Online**, by measuring actual user response
- **Offline**, by using existing data sets
  - **Prediction accuracy** with rating data (RMSE)
  - **Top-*N* accuracy** with ratings, purchases, clicks, etc. (IR metrics – MAP, MRR, P/R, AUC, nDCG)

Offline Evaluation



The Candidate Set



Often: $C_u = I \setminus R_u$
(all items not rated in training)

Recommender is a *classifier* separating relevant items (T_u) from decoy items (D_u)

Missing Data

☐ *Zootopia*

☑ *The Iron Giant*

☑ *Frozen*

☒ *Seven*

☐ *Tangled*

RR = 0.5

AP = 0.417

IR metrics assume a **fully coded corpus**

- Real data has unknowns
- Unknown = irrelevant

For recommender systems, this assumption is 👖🔥

Misclassified Decoys

☐ *Zootopia*
☑ *The Iron Giant*
☑ *Frozen*
☒ *Seven*
☐ *Tangled*

3 possibilities for *Zootopia*:

- I don't like it
- I do but data doesn't know
- I do but **I don't know yet**

RR = 0.5

AP = 0.417

Misclassified Decoys

If I would like *Zootopia*

But have not yet seen it

Then it is likely a **very good** recommendation

But the recommender is penalized

How can we fix this?

IR Solutions

Rank Effectiveness

- Only rank test items, don’t pick from big set
- Requires ratings or negative samples

Pooling

- Requires judges – doesn’t work for recsys

Relevance Inference

- Reduces to the recommendation problem
- Can we really use a recommender to evaluate a recommender?

Sturgeon’s Law

Ninety percent of everything is crud.

– T. Sturgeon (1958)

Only 1% is ‘really good’

– P. S. Miller (1960)

Sturgeon's Decoys

Most items are not relevant.

Corollary: a randomly-selected item is probably not relevant.

Random Decoys

- Generalization of One-Plus-Random protocol (Cremonesi et al. 2008)
- Candidate set contains
  - Test items
  - Randomly selected decoy items

One Plus Random tries to recommend each test item separately

How Many Decoys?

Koren (2008): right # is open problem, used 1000

Our origin story: find a good number or fraction

Modeling Goodness

Starting point: $\Pr[i \in G_u]$, probability i is good for u

goodness rate g

Want: $\Pr[D_u \cap G_u = \emptyset] \geq 1 - \alpha$

high likelihood of no misclassified decoys

Simplifying assumption: goodness is independent

$$\Pr[D_u \cap G_u = \emptyset] = \prod_{i \in D_u} \Pr[i \notin G_u] = (1-g)^N$$

What's the damage?

For $\alpha = 0.05$ (95% certainty), $N = 1000$

$$1 - g = 0.95^{\frac{1}{N}}$$
$$g = 0.0001$$

Only 1 in 10,000 can be relevant!

MovieLens users like 10s to 100s of 25K films

Why so serious?

If there is even one good item in the decoy set …

… then it is the recommender’s **job** to find that item

If no unknown items are good, why recommend?

Popularity Bias

Evaluation naively favors popular recommendations

Why?

Popular items are more likely to be rated

And therefore more likely to be ‘right’

Problem: how much of this is ‘real’?

Sturgeon and Popularity

Random items are …

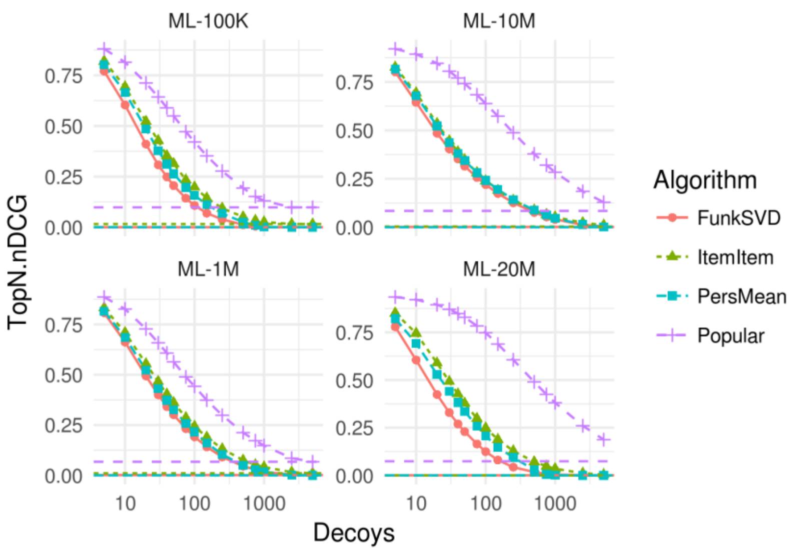
… less likely to be relevant (we hoped)

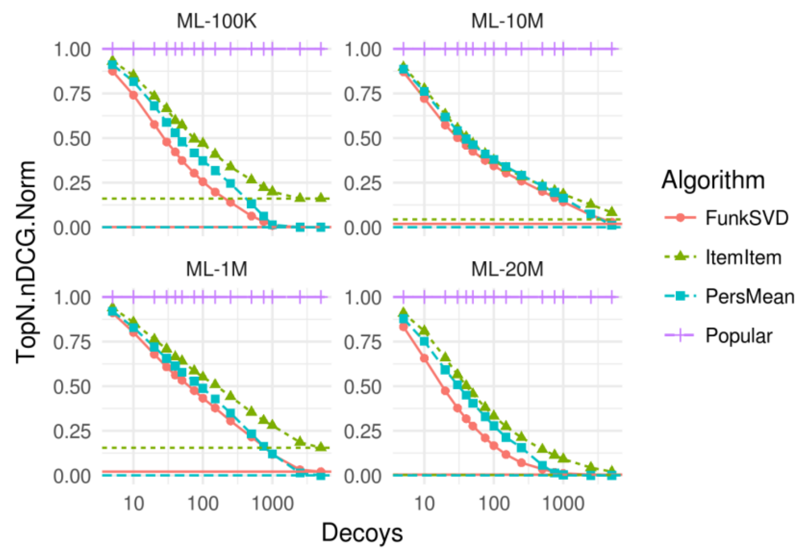
… less likely to be popular

Result: popularity is **even more** likely to separate test items from decoys

oops

Empirical Results





Empirical Findings

- Didn’t see theoretically-expected impact
- Absolute difference depends on decoy set size
  - Statistical significance depends on set size!
- No clear inflection points for choosing a size
- Algorithm ordering unaffected

Takeaways

Random decoys seem useful, but …

… have unquantified benefit

… may not achieve benefit

… have complex problems

… hurt reproducibility

Future Work

- Compare under Bellogin’s techniques
  - What happens w/ decoy sizes when neutralizing popularity bias?
- Try with more domains
- Try one-class classifier techniques
- Extend theoretical analysis to ‘Personalized Sturgeon’s Law’

Thank you

- Thanks to Sole Pera and the PIReTs
- Texas State for supporting initial work

Questions?

People and
Information
Research
Team

BOISE STATE
UNIVERSITY

https://goo.gl/bfVg1T