# RECOMMENDATIONS FOR THE RECOMMENDERS

## REFLECTIONS ON PRIORITIZING DIVERSITY IN THE RECSYS CHALLENGE

*Lucien Heitz*, Oana Inel* and Sanne Vrijenhoek^*

**Universität Zürich*

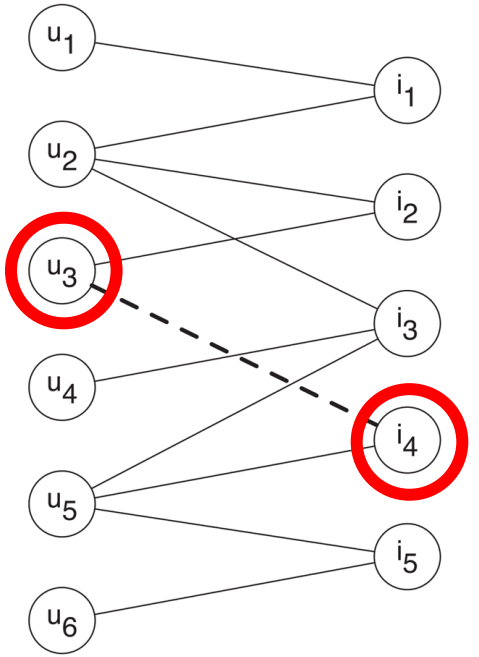
*^University of Amsterdam*

# INTRODUCTION & GOALS

- Focus on technical and normative challenges.
- Investigating how RS resonate with editorial values.
- Rank articles based on the user's personal preferences.
- Primary evaluation metric is AUC.

# CHALLENGE SUBMISSION

- Focus on **diversity** as beyond-accuracy objective
- Using RP3-β random walk:
  - Based only on UxI interactions
  - Diversity-accuracy trade-off
- Data pre-processing:
  - Focus on "meaningful" interactions
  - Give priority to new items



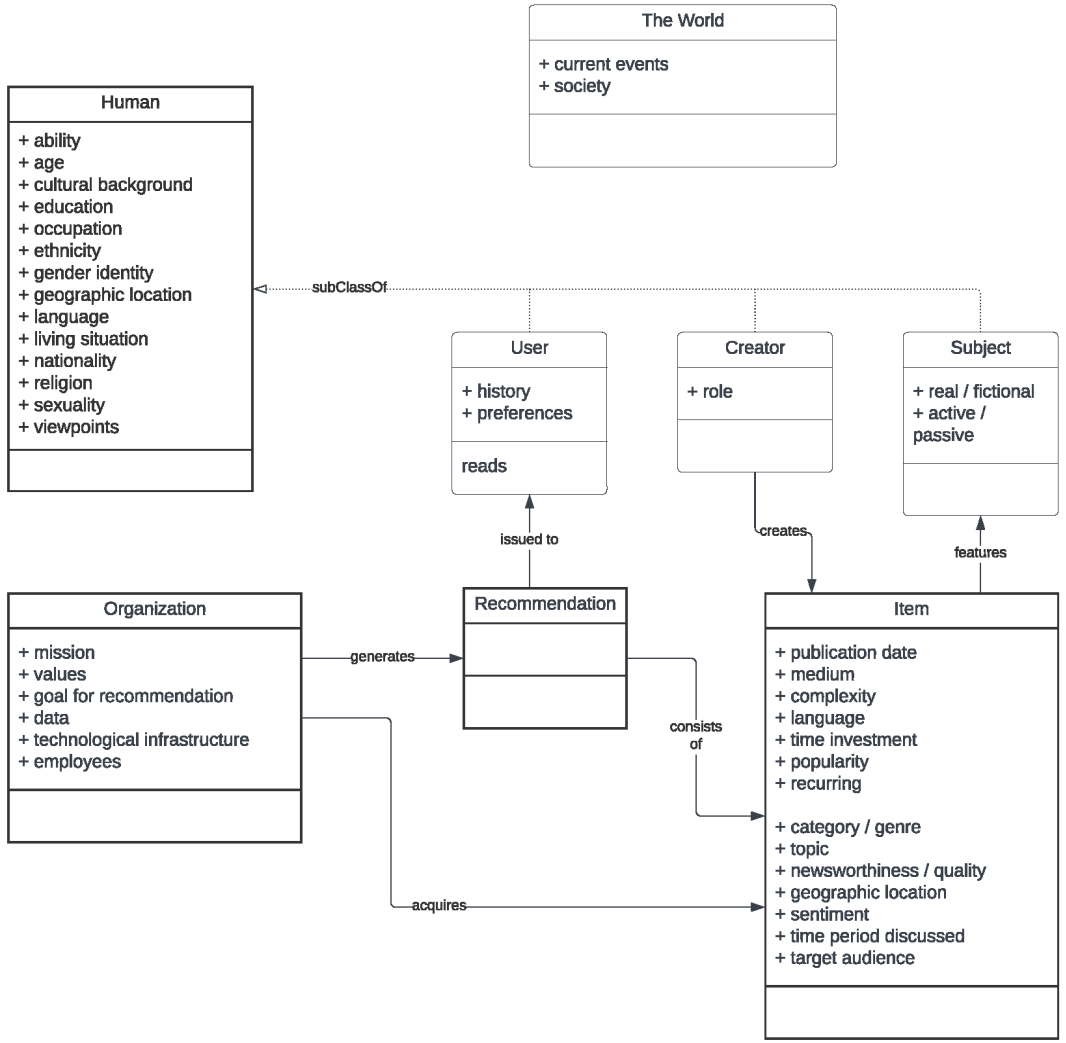
# RESULTS AND LIMITATIONS

- We see a mismatch between challenge goals and submissions
- We are unsure about:
  - Reflecting normative dimensions
  - Editorial values
- We identify two limitations:
  - Definition of diversity
  - Recommendation task

| Model | AUC | MRR | NDCG | ILD* | NOV* | SER* |
|---|---|---|---|---|---|---|
| RP3-β | 0.5005 | 0.3167 | 0.3505 | 0.7549 | 11.101 | 0.8069 |
| EB | 0.5684 | 0.3517 | 0.3951 | 0.8402 | 3.071 | 0.7915 |
| Top 3 | 0.7643 | 0.5501 | 0.6117 | 0.6255 | 5.771 | 0.7568 |

# DEFINING DIVERSITY

- Challenge focused on "RecSys standard" diversity: Intra-List Distance (ILD) of topics
- *"Personalized recommendations resulted in exposure to softer news content and a less diverse variation in the exposure of topics, actors, or issues."* (Einarsson et al., 2024)




Sanne Vrijenhoek, Savvina Daniil, Jorden Sandel, and Laura Hollink. 2024. Diversity of What? On the Different Conceptualizations of Diversity in Recommender Systems. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). Association for Computing Machinery, New York, NY, USA, 573–584. https://doi.org/10.1145/3630106.3658926

# DATASET STRUCTURE

- Many questions:
  - Which articles are included in the dataset? All published items? If so, how far does that go back? Or only the items users engaged with?
  - How was the candidate list generated?
  - How were the articles displayed on screen? Could there have been presentation bias through the UI? Did users actually want to read that article, or did they simply click the top item?
  - Should we really have replicated the MIND structure?
- It is very hard to account for diversity when a recommender can only 'choose' from on average 11 articles.

# RECOMMENDATION TASK

- Choose **one** item that will be recommended:
  - Diversity can only be considered at the aggregate level.
  - Does not reflect the requirements media organizations have for their recommendations.
- Probably a disproportionate focus on certain article categories.
- How would the newsroom rate the selection of articles?

If beyond accuracy objectives are indeed considered to be a priority, then this needs to be reflected accordingly in the evaluation procedure.

**Table 2.** Distribution of the different article categories (the whole dataset, what was in the users' reading history, the dataset after candidate selection, and what the user clicked), and the recommender approaches. For the recommendations the top 8 items are selected. The distribution shown does not account for ranking.

| | MIND | | | | Recommendations | |
|---|---|---|---|---|---|---|
| | all | candidate | history | clicked | LSTUR | NRMS |
| hard | 0,363 | 0,302 | 0,348 | 0,269 | 0,261 | 0,253 |
| soft | 0,636 | 0,698 | 0,622 | 0,730 | 0,739 | 0,747 |
| news | 0,305 | 0,233 | 0,279 | 0,235 | 0,215 | 0,224 |
| sports | 0,314 | 0,163 | 0,142 | 0,245 | 0,209 | 0,207 |
| finance | 0,058 | 0,069 | 0,069 | 0,034 | 0,046 | 0,029 |
| travel | 0,049 | 0,027 | 0,030 | 0,020 | 0,024 | 0,021 |
| video | 0,045 | 0,020 | 0,019 | 0,021 | 0,021 | 0,016 |
| lifestyle | 0,044 | 0,171 | 0,105 | 0,178 | 0,174 | 0,185 |
| foodanddrink | 0,043 | 0,068 | 0,050 | 0,057 | 0,078 | 0,077 |
| weather | 0,040 | 0,028 | 0,012 | 0,015 | 0,028 | 0,021 |
| autos | 0,030 | 0,030 | 0,036 | 0,024 | 0,019 | 0,019 |
| health | 0,028 | 0,034 | 0,047 | 0,034 | 0,047 | 0,041 |
| music | 0,013 | 0,044 | 0,027 | 0,035 | 0,042 | 0,057 |
| tv | 0,013 | 0,047 | 0,082 | 0,048 | 0,046 | 0,041 |
| entertainment | 0,008 | 0,029 | 0,034 | 0,024 | 0,029 | 0,034 |
| movies | 0,008 | 0,038 | 0,038 | 0,030 | 0,024 | 0,028 |

# WHY SHOULD YOU (AND RECSYS) CARE?

- *"it is not even clear if slightly higher accuracy values are relevant in terms of adding value for recommendation consumers or providers"* (Dacrema et al, 2019).
- Conferences have agenda-setting power too.
- There are enough technical challenges left to be solved besides accuracy.


A THANK-YOU
TO THE
ORGANIZERS

# RECOMMENDATIONS FOR THE RECOMMENDERS

- Incentivize submissions for beyond-accuracy objectives by
  1. providing clear guidelines on and explicitly defining the types of beyond-accuracy factors to be taken into account,
  2. providing the necessary contextual understanding on the subset of items that the dataset contains, and
  3. introducing multiple leaderboards, such that the evaluation of the submitted systems is more easily performed
- Promote and include online testing as a standard component of the RecSys challenge. We recommend online evaluations to be performed on the top submissions for all available leaderboards.
- Streamline the organization and evaluation of submissions by choosing a suitable, single environment to host and organize the challenges over the years.
- Promote forming inter- and cross-disciplinary participation in the challenge beyond designing, developing, and submitting recommender models.
- Encourage the challenge hosts to allow usage of the dataset for research purposes beyond the scope of the challenge submissions.