WHEN PREFERENCE IS NOT ENOUGH

Why Recommender Systems Require Human-Aware Evaluation for Children

Robin Ungruh, Alejandro Bellogín, Dominik Kowald, Sole Pera

PROBLEM

- Recommender systems **evaluation** conventionally relies on accuracy-based proxies for user preference.
- Children have **additional and unique needs** that are overlooked by traditional paradigms.
- **Content maturity** serves as a established proxy for assessing alignment with children's needs.

To what extent do recommender systems fit children's maturity levels?

EXPERIMENT SETUP

- **Data**: MovieLens-1m (ML) & MyAnimeList (MAL)
- **Evaluation Criteria**:
  - Traditional "Preference-centered" Evaluation
  - Percentage of Exceedances
  - Amplification/Suppression of Maturity Labels
- **Recommender**: *MostPop, ItemKNN, UserKNN, EASEr, RP3beta, Slim, MF2020, MultiVAE*

		ML		MAL	
		% items	*% interactions*	*% items*	*% interactions*
Not Rated	No classification	9.60	3.77	0.40	0.01
Approved	General audiences	9.17	7.04	-	-
G	All ages	5.01	6.01	11.97	2.16
PG	Parental guidance	16.13	21.98	6.50	2.16
PG-13	13+	16.23	16.32	52.67	57.65
R	17+	43.31	44.65	12.65	27.15
R+/NC-17	Adults only	0.48	0.11	10.32	10.35
Rx	Pornographic	0.02	0.00	5.49	0.49

Table 1: Dataset statistics.

FINDINGS

Traditional Performance

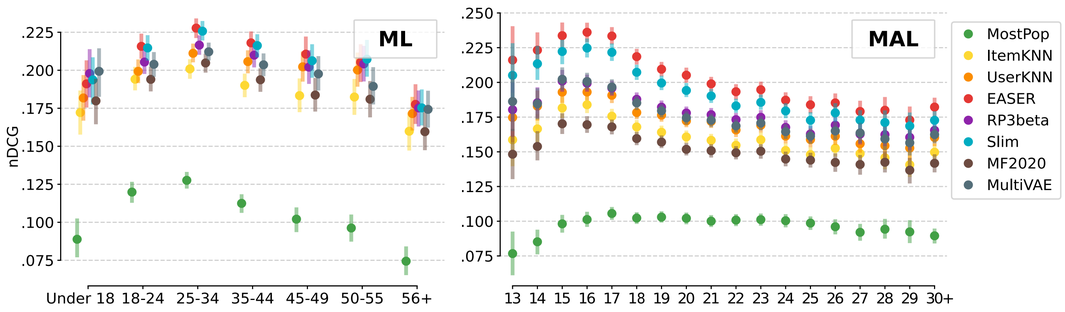

Figure 1: nDCG scores.

→Minor differences in performance.
Older teenagers not positioned as "underserved".

Maturity Consumption

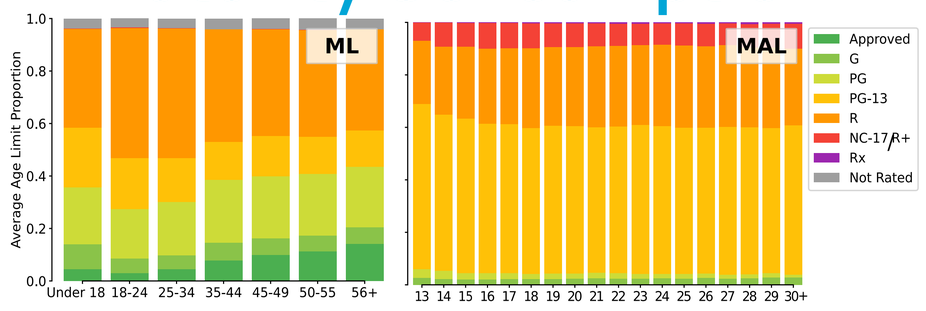

Figure 3: Proportion of maturity levels in user profiles.

→Children already consume "unsuitable" items.

High-Maturity Items

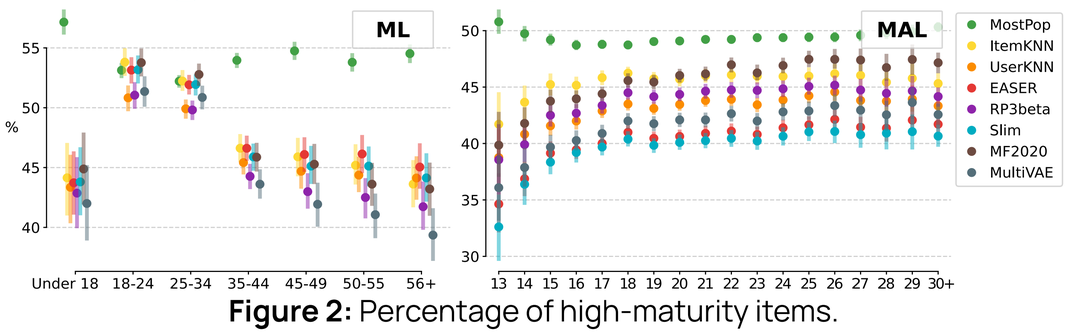

Figure 2: Percentage of high-maturity items.

→Items exceeding formal maturity levels are frequent.

Maturity Amplification

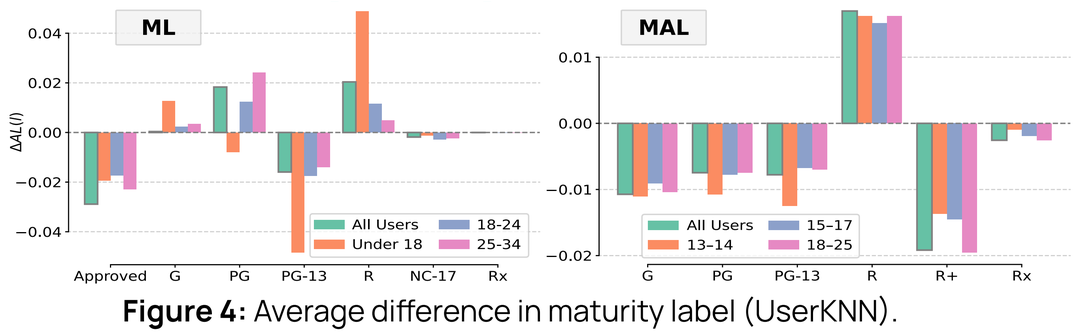

Figure 4: Average difference in maturity label (UserKNN).

→R-rated items are amplified while suitable items are suppressed in recommendations.

IMPLICATIONS

- Traditional paradigms provide an **incomplete picture** of recommendation quality for children.
- Recommenders expose and even **amplify exposure** to mature content.
- **Human-centric evaluation lenses** that explicitly capture hetero-geneous needs are necessary.

REFERENCES

Ungruh R, Bellogín, A, Kowald, D, Pera, M.S., 2025. Impacts of Mainstream-Driven Algorithms on Recommendations for Children Across Domains: A Reproducibility Study. In *RecSys 2025*.

Konstan, J., Terveen, L, 2021. Human-centered recommender systems:

Origins, advances, challenges, and opportunities. *AI Magazine 42, 3 (2021)*

Thompson, K.M., Yokota, F., 2004. Violence, sex, and profanity in films: Correlation of movie ratings with content. *Medscape General Medicine* 6, 3 (2004), 3.