Joint Evaluation of Fairness and Relevance in Recommender Systems with Pareto Frontier

PAPER CODE

Theresia Veronika Rampisela[1] **Tuukka Ruotsalo**[1, 2] **Maria Maistro**[1] **Christina Lioma**[1]

[1]University of Copenhagen, Denmark [2]LUT University, Finland

Email: {thra, tr, mm, c.lioma}@di.ku.dk 𝕏: @theresia_v_r

Distance to Pareto Frontier (DPFR):
How close are the models to an ideal balance of fairness and relevance?

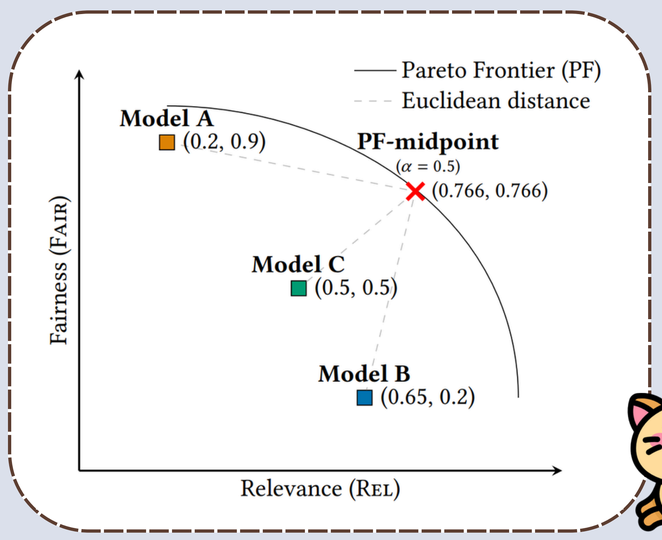
We propose DPFR, a Pareto-optimal-based evaluation approach to measure recommender system fairness & relevance jointly.

Recommender systems: systems that can **match/recommend items to users**, such that the users will find the recommended items **relevant**
Relevance: an item is relevant to a user if the user **likes** it or finds it **useful**
Fairness: broadly defined as **treating users/items equally**

We focus on individual item fairness: ensures that each item is recommended a similar amount of times across all users

Goal: find the **most balanced** model in terms of both fairness and relevance

Solution: take the **distance** between the model scores and the midpoint of the Pareto Frontier



What the Pareto Frontier means:
Given a **certain level of relevance**, what is the **maximum achievable fairness** based on the dataset composition?

"**Model A** is the **fairest**, **Model B** has the **highest relevance**, **Model C** is the **closest to the Pareto Frontier,** so it is the most balanced!"

1 Background

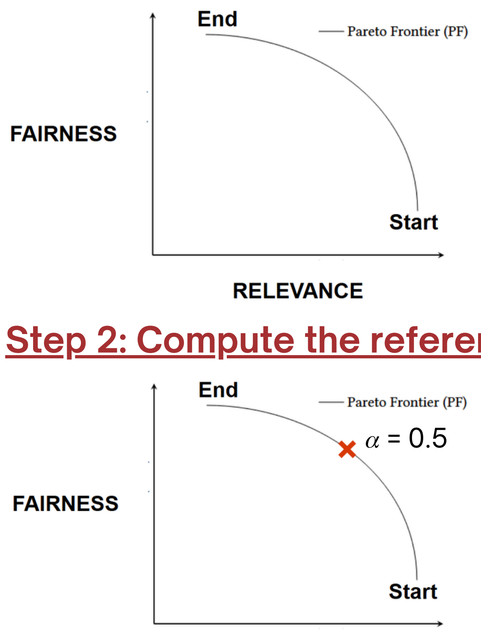
- Fairness and relevance are **two important aspects** of recommender systems (RSs).
- Typically, they are evaluated either:
  (i) **separately** by individual measures of fairness and relevance
  (ii) **jointly** with a measure that accounts for fairness w.r.t. relevance

However,
Type (i) measures **do not provide a reliable joint estimate** of the model relevance and fairness
Type (ii) measures **do not empirically account for both aspects** well

Motivated by this, we present a **new approach** for jointly evaluating fairness and relevance in RSs: **Distance to Pareto Frontier (DPFR)**.

2 Distance to Pareto Frontier

Step 1: Generate the Pareto Frontier (PF)



Start: use the test set to **create maximally relevant recommendation**

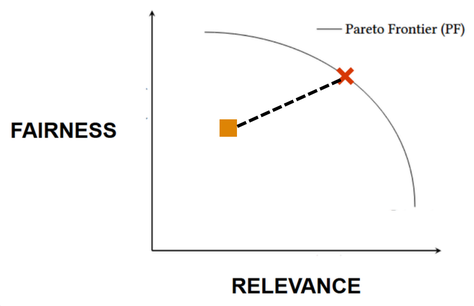
Iteratively replace items to **increase fairness**

End: fairest possible recommendation

Step 2: Compute the reference point

Select a point in the PF based on α.
α controls the relative position between the start & end points.
- α = 0 only considers relevance
- α = 1 only considers fairness

Step 3: Compute distance from the model to the reference point



The distance between the model and the reference point is the **DPFR score**

DPFR is **modular, tractable,** and **intuitive**. It can be used with **existing measures for relevance and fairness**, and **allows for different trade-offs** of relevance and fairness.

3 Experimental Setup

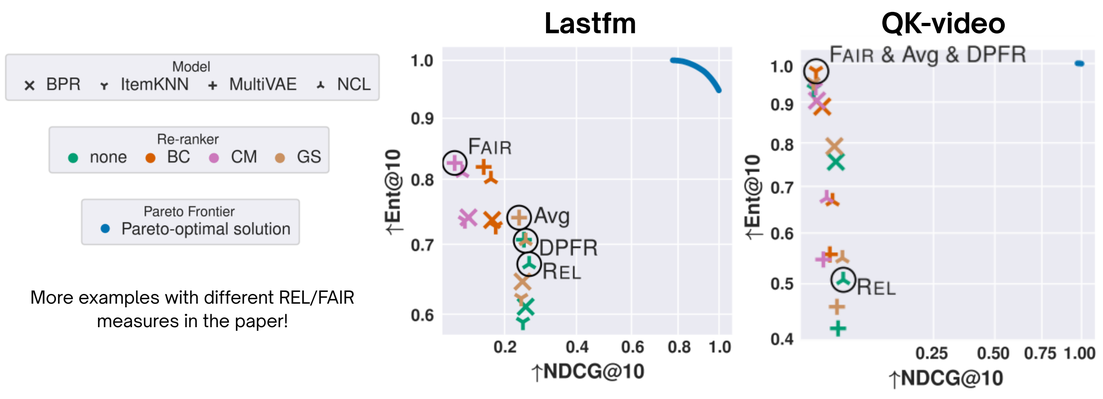
Data
- **6 interaction datasets**: Lastfm (music), Amazon-lb (e-commerce), QK-video, Jester (jokes), ML-10M & ML-20M (movies)

Models
- **4 recommenders:** ItemKNN, BPR, MultiVAE, NCL
- **3 fair rerankers:** Greedy Substitution (GS), COMBMNZ (CM), Borda Count (BC)

Evaluation
- **Single-aspect** measures: 6 relevance (REL) + 5 fairness (FAIR)
- **Joint** measures of relevance & fairness:
  - 5 joint measures of relevance and fairness
  - Avg: : Averaging relevance + fairness score
  - DPFR: Distance to Pareto Frontier

4 Our Findings

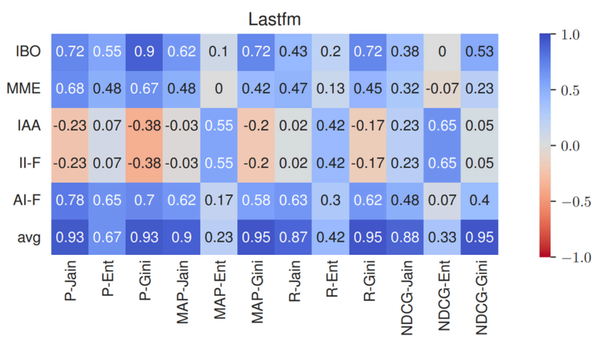
Finding #1
The best model based on DPFR always differs from the best model based on relevance. For fairness, it differs half the time.



More examples with different REL/FAIR measures in the paper!

Finding #2
Existing joint evaluation measures are not a reliable proxy for DPFR

Kendall's Tau correlation (τ) of joint measures (y-axis) and DPFR (x-axis).
τ ⩾0.9 means the joint measures rank models equivalently to DPFR.



Finding #3
The best model based on **Avg differs from DPFR** up to **83%** of the time

	Set-based	Rank-based	All
Lastfm	50.00	66.67	58.33
Amazon-lb	0.00	0.00	0.00
QK-video	16.67	0.00	8.33
Jester	16.67	83.33	50.00
ML-10M	0.00	66.67	33.33
ML-20M	0.00	50.00	25.00
All datasets	13.89	44.44	29.17

This work is supported by: ADD algorithms data & democracy

Presented at The ACM Web Conference 2025 THE WEB CONFERENCE ACM

Credits: Icons by Freepik - Flaticon, graphics by Canva assets