Alleviating the unfairness issue with knowledge-aware recommendation models

EEMCS, Delft University of Technology, The Netherlands

Author: Yoan Popov | *Y.Z.Popov@student.tudelft.nl*

Supervisor: Masoud Mansoury

Background

Recommender systems help users navigate large search spaces by offering personalized suggestions. Traditional approaches include content-based, collaborative, and knowledge-based methods, and modern systems often combine these to form hybrid models for improved performance.

Knowledge-aware models leverage structured side information (e.g., knowledge graphs) to enhance recommendation quality. These graphs capture semantic relationships among entities, which improves precision, and helps mitigate cold-start and data sparsity issues.

Hybrid systems combine two or more recommendation techniques to leverage their strengths and offset weaknesses. Hybridization strategies include weighted models, feature combination, and cascading —many of which are used in recent deep recommender systems [1]. Hybrid systems are oftentimes also knowledge-aware.

Fairness refers to the ethical evaluation of how recommendation systems impact different user or item groups. Current work focuses heavily on group-level fairness (e.g., gender or ethnicity) and mostly uses static, outcome-based fairness metrics due to ease of measurability [2].

Knowledge Gaps: Although knowledge-aware and hybrid models are widely studied for their accuracy, their fairness performance is underexplored. Most research targets optimization of predictive power [3], leaving a gap in understanding how these models behave under fairness criteria, which this study aims to address.

Research Question

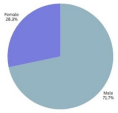
Can knowledge-aware recommendation models alleviate the unfairness issue, and if so, to what degree?

RQ1: Do knowledge-aware models perform better than other paradigms on metrics related to fairness?
RQ2: Does adjusting the relative weights of components in the loss function of a knowledge-aware recommender system lead to improved performance on fairness and accuracy metrics?
RQ3: Given the findings in RQ2, can optimizing for fairness, based on the selected metrics, affect the accuracy of the models in question? If so, to what extent?

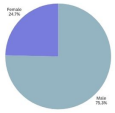
Methodology

Dataset:
The MovieLens 1M dataset was utilized for this study . Knowledge-graph was acquired via RecSysDatasets. Duplicate removal and 5-core filtering was applied.

Filtered users pie chart



Filtered interactions pie chart

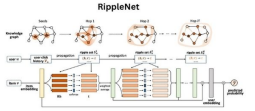


Models:

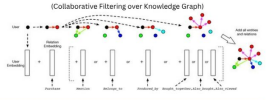
Knowledge-aware models

CKE
(Collaborative Knowledge Base Embedding)

RippleNet



CFKG
(Collaborative Filtering over Knowledge Graph)



General Recommender Baselines

POP (Popular) - recommends the most popular items (i.e., those with the highest number of interactions) to all users

Random - recommends items randomly, without considering user preferences or item popularity.

ItemKNN - collaborative filtering model computes similarities between items and recommends items similar to those a user has interacted with

Fairness-aware model

PFCN-PMF - extends Probabilistic Matrix Factorization by inserting an adversarial "filter" module that removes user-sensitive at tributes (e.g., gender, age) from learned embeddings before scoring.

Metrics:

Accuracy

Evaluate how well the system retrieves and ranks relevant items within its top recommendations, considering both the number of hits and their positions. Overall discrimination between relevant and irrelevant items across users is also assessed

(notable metrics: Recall, NDCG)

User-side Fairness

They quantify how prediction errors differ between demographic groups, both in direction (whether one group's ratings are systematically over- or under-estimated) and in magnitude. Additionally, they measure group-level parity by comparing average predicted scores

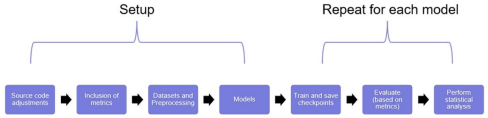
(notable metrics: Value unfairness family)

Item-side fairness

Calculate percentage of less-frequently chosen versus highly popular content in the final recommendation list. Diversity is further captured by measuring inequality in item exposure, where more equal distributions indicate broader coverage of items.

(metrics: Popularity Percentage, Tail Percentage, Gini Index)

Pipeline and Experimental setup

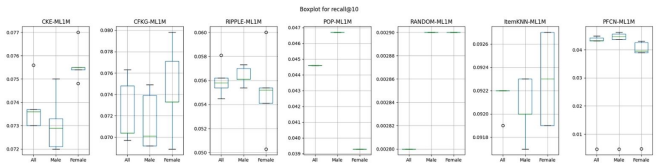


Framework: RecBole 1.2.1, combined with FairRec (RecBole Fork)

Parallel CPU execution and logging governed by Weights & Biases

Results

RQ1



:Boxplots for metric Recall@10, grouped by gender. The plots are based on the top 5 best-performing runs per model for this metric

The boxplots suggest that model performance varies more for the female group. However, it is possible, especially for knowledge-aware models, to choose a hyperparameter set that achieves better performance for that group.

Accuracy Metrics

Model	Z-score
ItemKNN	1.205
CKE	0.617
CFKG	0.544
RippleNet	0.183
PFCN	-0.203
POP	-0.209
Random	-2.137

User Fairness Metrics

Model	Z-score
POP	0.447
Random	0.439
CFKG	0.123
PFCN	0.078
CKE	-0.048
RippleNet	-0.444
ItemKNN	-0.595

Item Fairness Metrics

Model	Z-score
Random	2.433
CFKG	-0.242
CKE	-0.332
RippleNet	-0.374
ItemKNN	-0.442
PFCN	-0.520
POP	-0.522

Z-score normalization was applied on the selected metrics. The means for each model's z-score were calculated per row, inverting the z-score result wherever necessary, such that the final aggregated score implies "higher is better".

Takeaway: No single model or paradigm emerged as a universal champion across all performance domains. While the collaborative-filtering model ItemKNN dominated pure accuracy metrics, its performance on user-side fairness was notably poor. Conversely, simpler baselines like POP and Random excelled in specific fairness domains – POP in user-side fairness and Random in item-side diversity– albeit at the cost of accuracy. Knowledge-aware models occupied an interesting middle ground. CFKG demonstrated the best overall user-side fairness among similar models and even surpassed the fairness-aware PFCN-PMF model in this aggregate category. CKE, while strong in aggregate accuracy, did not particularly stand out for metrics from any fairness side, but was not a poor performer either. RippleNet offered competitive item-side diversity, but user-side fairness was not its strong suit.

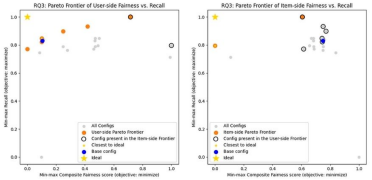
RQ2

Takeaway 1: The investigation into loss component weighting (RQ2) for CKE and RippleNet revealed the significant leverage provided by the knowledge graph component weight, while the recommendation loss weight showed minimal impact within the tested range

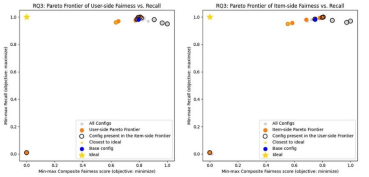
Takeaway 2: Optimizing for one fairness aspect or user group via the knowledge graph component may inadvertently affect others, necessitating careful, context-specific tuning.

Takeaway 3: It may possible to further develop the loss functions of those models by making the separate loss component weights learnable, adjusting them based on internally calculated fairness and accuracy metrics.

RQ3



RippleNet Pareto Frontiers



CKE Pareto Frontiers

Takeaway: The trade-off is not always severe - carefully selected hyperparameter configurations can lead to models that are both reasonably accurate and demonstrably fairer. It is also worth noting that there seems to be a tradeoff between user side fairness and item-side fairness. While there was some overlap between the frontiers, the majority of the points, which belonged to one fairness domain's frontier, did not belong to the frontier of the other fairness domain.

Model	Avg. user-side fairness change	Recall@10 change
CKE	17.13%	-2.04%
RippleNet	-0.03%	0.36%

Model	Avg. item-side fairness change	Recall@10 change
CKE	18.28%	-3.12%
RippleNet	80.63%	-0.72%

Conclusion & Future work

1: No single model universally excels across accuracy and all fairness dimensions. Knowledge-aware models demonstrate a promising ability to achieve a more holistic performance, often ranking competitively across accuracy, user-side, and item-side fairness domains without necessarily dominating any single one.

2: The weighting of the knowledge graph component proved to be a significant lever for tuning, capable of substantially impacting both accuracy and various fairness metrics

3: The exploration of the fairness-accuracy trade off indicated that substantial gains in specific fairness dimension might be achievable with only minor compromises in accuracy. A trade-off between user-side and item-side fairness was also displayed, implying a more complex relationship between those domains

Future work

- Broader evaluations with more datasets, diverse group definitions, more model types, and multiple splits would enhance the robustness of the findings.
- More well-defined fairness metrics need to be calculated across a larger component-weight search space
- The trade-off seen between user-side and item-side fairness needs to be studied further.
- More work needs to be done in order to ensure easier and more reproducible pipeline setups when using such open-source projects.

References

[1] Iana A, Alam M, Paulheim H. A survey on knowledge-aware news recommender systems. Semantic Web. 2022;15(1):21-82. doi:10.3233/SW-222991

[2] Deldjoo, Y., Jannach, D., Bellogín, A. et al. Fairness in recommender systems: research landscape and future directions. User Model User-Adap Inter 34, 59–108 (2024). https://doi-org.tudelft.idm.oclc.org/10.1007/s11257-023-09364-z

[3] Çano E, Morisio M. Hybrid recommender systems: A systematic literature review. Intelligent Data Analysis. 2017;21(6):1487-1524. doi:10.3233/IDA-163209