Exploring the Effect of Context-Awareness and Popularity Calibration on Popularity Bias in POI Recommendations

Andrea Forster, Simone Kopeinik, Denis Helic, Stefan Thalmann, and Dominik Kowald
TU Graz, University of Graz & Know Center Research GmbH (Graz, Austria)

Introduction

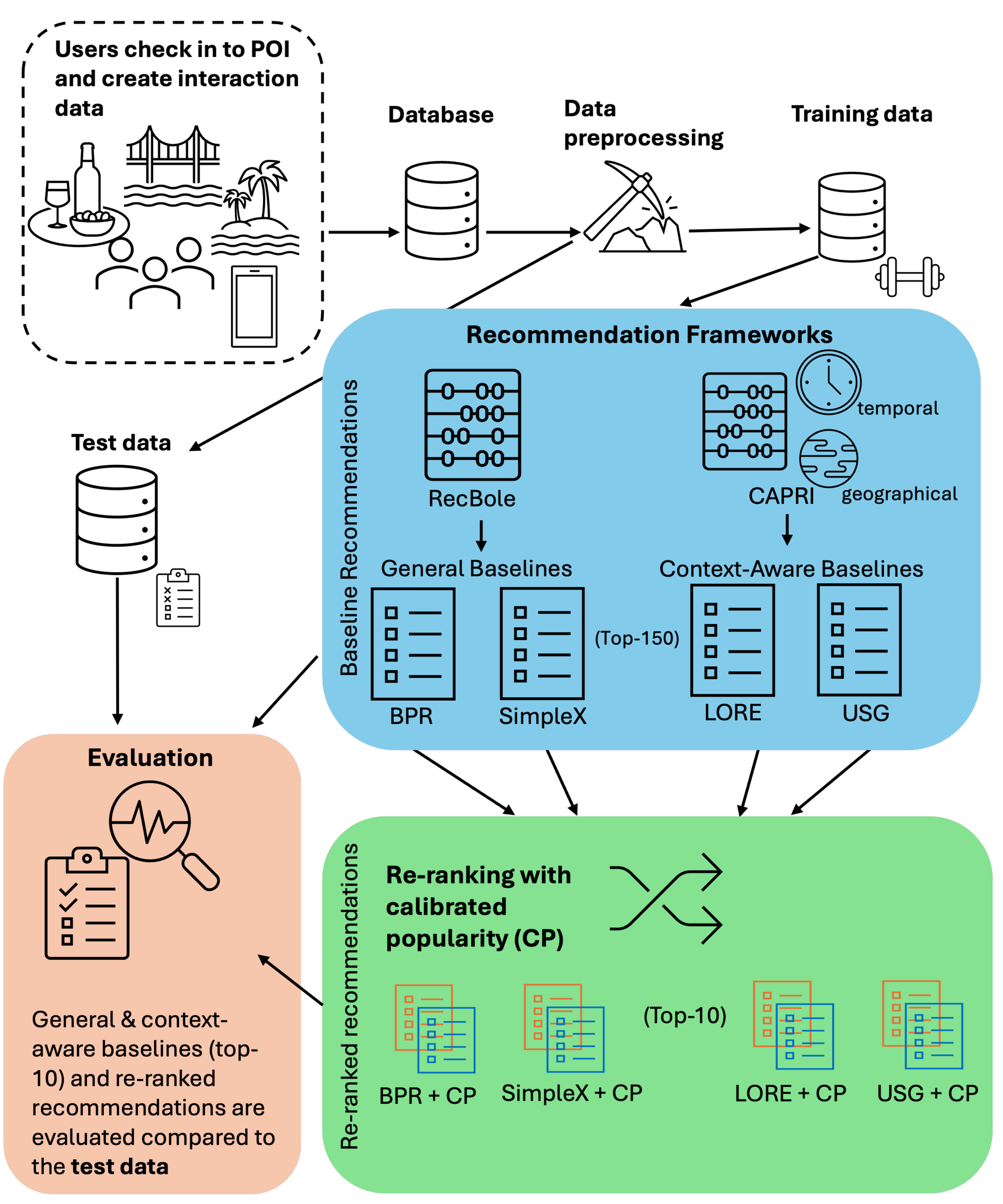
- Point-of-interest (POI) recommender systems can be compromised by **popularity bias**, disadvantaging niche users and less popular, yet potentially meaningful places.
- This paper provides empirical insights into the **individual** and **combined** effects of **context-aware models** and **calibrated popularity** on recommendation accuracy and popularity bias.

Datasets

Dataset	Users	Items	Check-ins	Sparsity
Brightkite	600	794	15,341	0.967798
Foursquare	1,500	2,804	69,401	0.983500
Gowalla	1,500	7,579	53,679	0.995278
Yelp	1,500	4,515	35,288	0.994790

Methods

- **Context-aware POI recommendations**: Approaches that consider social, geographical, temporal, categorical, and/or sequential influences.
- **Calibrated popularity (CP)**: Re-ranking technique balancing the popularity distribution in the user profiles and recommendation lists (CP_H tuned for accuracy and $CP_{\Im}$ tuned for bias mitigation)



Link to Paper

Results

RQ1: To what extent can context-aware recommendations (LORE, USG) and calibration-based debiasing (CP) individually mitigate popularity bias in POI recommendations, and how does this impact accuracy, compared to a non-contextual baseline (BPR)?

Symbols indicate the preferred direction for each metric; best values shown in bold. For BPR, absolute values are shown; Δ% values for LORE, USG, CP_H and $CP_{\Im}$. Significant relations are indicated by ** via t-test ($p < 0.05$), Bonferroni-corrected for each metric.

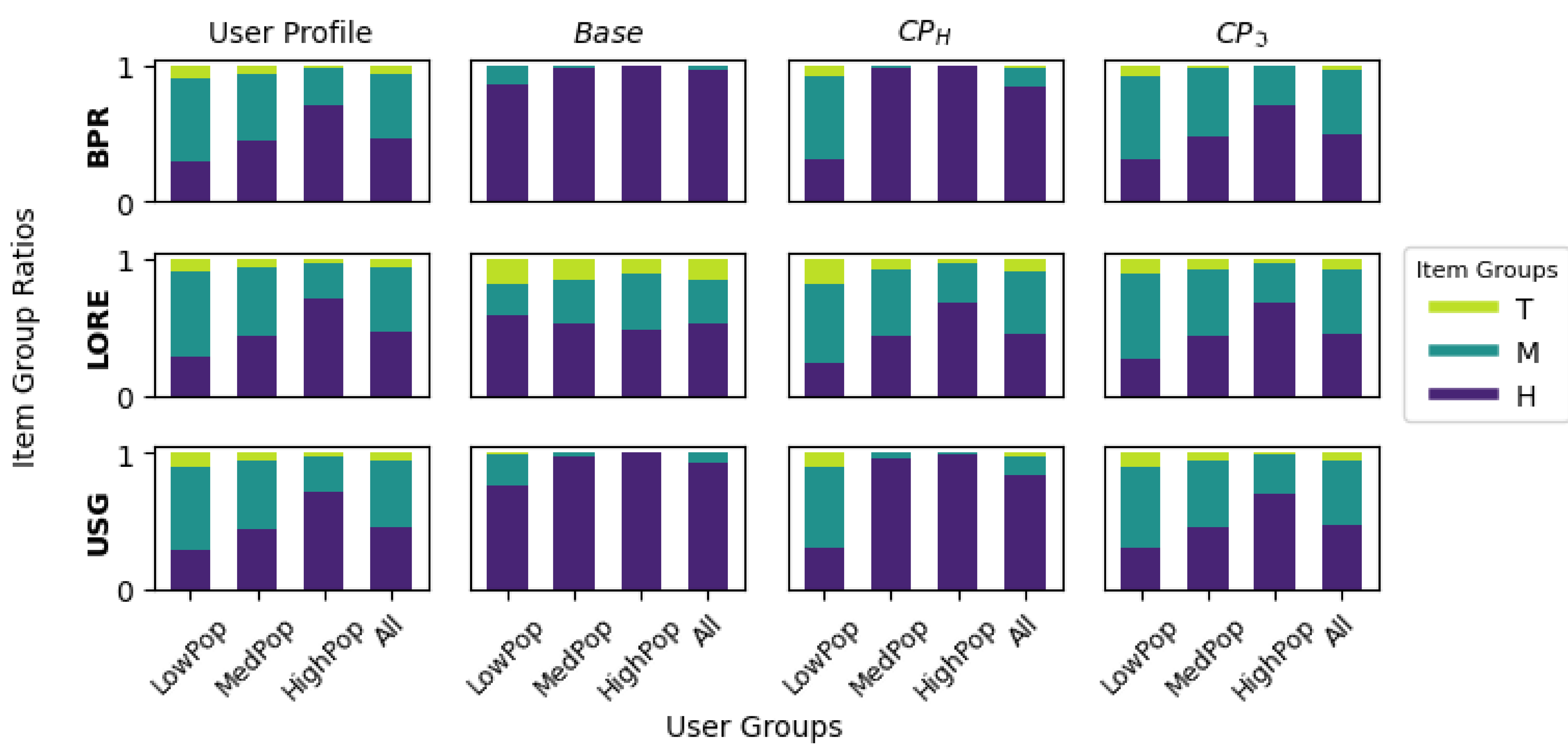
Group	nDCG ↑ ∣ Δ% nDCG					ARP ↓ ∣ Δ% ARP					PopLift → 0 ∣ Δ% PopLift				
	BPR	LORE	USG	CP_H	$CP_{\Im}$	BPR	LORE	USG	CP_H	$CP_{\Im}$	BPR	LORE	USG	CP_H	$CP_{\Im}$
Foursquare															
LowPop	0.0395	-56.62%**	-43.28%**	+0.54%	-21.28%	0.0795	-91.30%**	+94.71%**	-0.08%	-30.97%**	4.3299	-110.89%**	+149.05%**	-0.11%	-39.98%**
MedPop	0.1084	-88.77%**	-11.81%**	-0.04%	-14.92%**	0.1009	-93.69%**	+32.79%**	-0.41%**	-17.69%**	2.2977	-134.47%**	+48.68%**	-0.72%**	-26.28%**
HighPop	0.1655	-96.63%**	-2.11%	-0.14%	-6.43%**	0.1089	-94.31%**	+18.09%**	-0.37%**	-6.85%**	1.2302	-171.04%**	+33.50%**	-0.82%**	-13.57%**
All	0.1060	-88.83%**	-11.13%**	-0.03%	-12.74%**	0.0982	-93.44%**	+39.55%**	-0.35%**	-17.44%**	2.4906	-129.88%**	+82.08%**	-0.52%**	-29.79%**
Yelp															
LowPop	0.0192	+126.17%**	+12.65%	+6.94%	+6.94%	0.0040	-63.60%**	-15.61%**	-38.46%**	-38.46%**	1.7702	-101.15%**	-25.27%**	-65.20%**	-65.20%**
MedPop	0.0304	-33.66%**	+2.22%	+1.15%	-27.81%**	0.0079	-75.92%**	-5.00%**	-0.14%**	-35.00%**	2.3983	-107.50%**	-7.77%**	-0.22%**	-50.67%**
HighPop	0.0650	-75.53%**	-0.94%	+0.00%	-17.65%**	0.0093	-74.34%**	+0.54%	-0.01%	-18.58%**	1.0542	-145.19%**	+1.13%	-0.04%	-37.78%**
All	0.0350	-31.70%**	+2.19%	+1.36%	-20.24%**	0.0074	-74.20%**	-4.74%**	-4.21%**	-31.24%**	2.0039	-110.34%**	-9.93%**	-11.68%**	-51.88%**

RQ2: Does the combination of context-aware POI recommendations and calibration-based debiasing (CP) improve the trade-off between recommendation accuracy and popularity bias, compared to their respective purely context-aware versions (LORE, USG)?

Model	Group	nDCG ↑ (Δ% nDCG)			ARP ↓ (Δ% ARP)			PopLift → 0 (Δ% PopLift)		
		Base	CP_H	$CP_{\Im}$	*Base*	CP_H	$CP_{\Im}$	*Base*	CP_H	$CP_{\Im}$
Foursquare										
LORE	*LowPop*	0.0172	0.0255 (+48.51%)	0.0255 (+48.51%)	0.0069	0.0119 (+72.26%**)	0.0119 (+72.26%**)	-0.4715	-0.1371 (+70.92%**)	-0.1371 (+70.92%**)
LORE	*MedPop*	0.0122	0.0212 (+74.07%**)	0.0212 (+74.07%**)	0.0064	0.0132 (+107.75%**)	0.0132 (+107.75%**)	-0.7920	-0.5712 (+27.88%**)	-0.5712 (+27.88%**)
LORE	*HighPop*	0.0056	0.0200 (+258.16%**)	0.0200 (+258.16%**)	0.0062	0.0149 (+139.93%**)	0.0149 (+139.93%**)	-0.8740	-0.6986 (+20.06%**)	-0.6986 (+20.06%**)
LORE	*All*	0.0118	0.0218 (+83.99%**)	0.0218 (+83.99%**)	0.0064	0.0133 (+106.32%**)	0.0133 (+106.32%**)	-0.7443	-0.5099 (+31.50%**)	-0.5099 (+31.50%**)
USG	*LowPop*	0.0224	0.0228 (+1.72%)	0.0222 (-1.03%)	0.1548	0.1353 (-12.62%**)	0.1247 (-19.47%**)	10.7837	9.0601 (-15.98%**)	7.9484 (-26.29%**)
USG	*MedPop*	0.0956	0.0937 (-2.03%)	0.0912 (-4.58%)	0.1340	0.1287 (-3.90%**)	0.1189 (-11.26%**)	3.4161	3.2405 (-5.14%)	2.9022 (-15.04%**)
USG	*HighPop*	0.1620	0.1602 (-1.12%)	0.1552 (-4.15%)	0.1286	0.1253 (-2.56%)	0.1168 (-9.19%**)	1.6424	1.5728 (-4.24%)	1.3966 (-14.97%**)
USG	*All*	0.0942	0.0928 (-1.54%)	0.0902 (-4.27%)	0.1371	0.1294 (-5.62%**)	0.1196 (-12.72%**)	4.5349	4.0709 (-10.23%**)	3.6103 (-20.39%**)
Yelp										
LORE	*LowPop*	0.0434	0.0426 (-1.69%)	0.0390 (-10.03%)	0.0014	0.0014 (+0.00%)	0.0015 (+5.37%)	-0.0203	-0.0210 (-3.66%)	0.0052 (+125.79%)
LORE	*MedPop*	0.0201	0.0254 (+26.22%)	0.0254 (+26.22%)	0.0019	0.0022 (+16.65%**)	0.0022 (+16.65%**)	-0.1798	-0.0473 (+73.71%**)	-0.0473 (+73.71%**)
LORE	*HighPop*	0.0159	0.0269 (+68.99%)	0.0269 (+68.99%)	0.0024	0.0032 (+34.00%**)	0.0032 (+34.00%**)	-0.4764	-0.3043 (-36.12%**)	-0.3043 (-36.12%**)
LORE	*All*	0.0239	0.0292 (+21.79%)	0.0284 (+18.77%)	0.0019	0.0023 (+18.46%**)	0.0023 (+19.29%**)	-0.2072	-0.0934 (+54.91%**)	-0.0882 (+57.44%**)
USG	*LowPop*	0.0216	0.0246 (+14.09%)	0.0246 (+14.09%)	0.0033	0.0022 (-35.61%**)	0.0022 (-35.61%**)	1.3230	0.4271 (-67.72%**)	0.4271 (-67.72%**)
USG	*MedPop*	0.0310	0.0315 (+1.58%)	0.0274 (-11.89%)	0.0075	0.0075 (+0.00%)	0.0048 (-35.75%**)	2.2118	2.2066 (-0.23%)	1.0415 (-52.91%**)
USG	*HighPop*	0.0644	0.0644 (+0.00%)	0.0496 (-22.87%)	0.0094	0.0093 (-0.26%)	0.0075 (-19.45%**)	1.0661	1.0605 (-0.53%)	0.6456 (-39.44%**)
USG	*All*	0.0358	0.0367 (+2.52%)	0.0313 (-12.70%)	0.0070	0.0068 (-3.55%)	0.0048 (-31.41%**)	1.8049	1.6215 (-10.16%**)	0.8395 (-53.49%**)

Item Popularity Group Distributions

Distribution of item groups (T, M, H) for the user groups (LowPop, MedPop, HighPop) in their user profile vs. BPR, LORE and USG *Base*, and the respective CP_H, and $CP_{\Im}$ in **Yelp**.



Conclusions

- When applied **individually**, the effectiveness of context-awareness varies between models, while CP fails to include sufficient T-items, even when tuned for bias mitigation ($CP_{\Im}$). Moreover, its effectiveness varies for different user groups when tuned for accuracy (CP_H).
- **Combining** context-awareness and CP can **balance individual trade-offs**: For **bias mitigation**, **LORE + CP** show best results. For **accuracy-focused outcomes** with some bias reduction, **BPR/USG + CP** (tuned for accuracy) show best results.
- **Future work:** Investigate the methods through user studies focusing on satisfaction and perceived quality across user groups.
- **GitHub**: `https://github.com/andreafooo/POI_RS_PopBias_Mitigation`