Supplementary information: Leveraging Monte Carlo Tree Search for Group Recommendation

Anonymous Author(s)

ABSTRACT

This document includes additional methodology details, Tables with results and charts. The companion repository can be found at: https://bit.ly/3WDr9fS.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

recommender systems, group recommendation, Large Language Models, heuristic search, novelty, diversity, fairness

ACM Reference Format:

Anonymous Author(s). 2024. *Supplementary information:* Leveraging Monte Carlo Tree Search for Group Recommendation. In . ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 EXPERIMENTAL SETUP

Data collection. Our evaluation used the *MovieLens* dataset[1], including user ids, movies, ratings, genres, and timestamps. No group information was originally included in the dataset. To account for shifts in user interests due to extensive watching histories, we only considered ratings created after January 1, 2016. We kept users rating more than 30 movies, resulting in over 29k users and over 56k movies. Assuming that the most recently watched movies better reflect current interests, enabling more timely recommendations, users' ratings were split based on timestamps, with the first 60% ratings as training, the next 20% as validation and the remaining 20% as testing. The test set of each group member only includes movies that were not part of any member train to avoid recommending seen elements again as part of the group recommendation [2]. Movies were considered liked if rated 4 or higher.

Group formation. To explore how fairness varies according to group size and the homogeneity/divergence of interests, in addition to the strategy described and reported in the main document, we adopted two additional strategies from prior research [2, 12] to create groups of 2 to 8 members:

i) **random**: simulating groups with unrelated members, we randomly selected users without replacement;

ii) **divergent**: Similar to the similar group creation strategy, but with one user intentionally dissimilar to the rest. We generated 1000 groups of each type. The average group similarity was 0.14, 0.23 and 0.15 for random, similar, and divergent groups, respectively. On average, every pair of users in the similar groups, rated a minimum of 9 (median 22) common movies, which, according to [1] helps to ensure the reliability of the similarity.

LLM selection. The LLM-based strategies used `Gemma:2b-instruct`, which showed a good balance between execution time and answer quality (in terms of prompt adherence). We also evaluated `Mistral:7b` and `Llama3`, but discarded them due to either excessive execution times on the available hardware or inappropriate prompt adherence, such as including Python code in the responses. Finally, given the exploratory nature of the work, we discarded `GPT` due to costs. We plan to expand the set of used LLMs in future works.

Baselines. We compared the performance of the proposed aggregation technique with different approaches: i) Commonly-used aggregation strategies based on social choice [4]: *additive*, *least-misery*, and *Borda* rank aggregation; ii) *GreedyLM* [12], a greedy technique using Least Misery Fairness; iii) *GFAR* [2][2] ; iv) *EP-FuzzDA* [3][3] ; and v) LLM-based aggregation (*LLM-agg*). The LLM was provided with groups' common watching history, users' individual watching history, the list of movies recommended to each user, and the task to solve. The LLM was asked to consider the provided information and, aggregate the individual recommended lists into one that satisfying all group members.

Evaluation metrics. Evaluating recommenders solely on the relevance of recommendations offers a partial view of their performance [7]. Therefore, we considered both relevance and different beyond-accuracy perspectives, namely: coverage, ordering and diversity/novelty metrics. For *relevance*, we selected precision and nDCG[4]. For *coverage*, we used *zero-recall* (as used in [2]), which measures the fraction of group members who did not receive any relevant recommendation. This metric can be also considered a fairness metric, as values close to zero indicate that every group member received at least one relevant item. As users watch movies in a certain order, and *order* is important in generating the sequences of candidate movies (by the MCTS algorithm), we considered Kendall's Tau rank correlation between the actual correlation order and the recommendation sequence[5]. For *diversity* and *novelty*, we followed the same definitions as in the reward strategies. Enhancing diversity/novelty is also associated with increased fairness and reduced biases [8], as they help broadening users' social experiences [8].

[1] https://grouplens.org/datasets/movielens/25m/

[2] https://github.com/mesutkaya/recsys2020

[3] https://github.com/LadislavMalecek/UMAP2021

[4] We followed the binary definition, which defines items as relevant or non-relevant.

[5] As correlation metrics require sequences to at least have the same number of elements, from the recommended movies, the analysis only included the sub-set of correctly recommended ones.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

	precision (↑)	nDCG (↑)	zrecall (↓)	corr	diversity (↑)	novelty (↑)	novelty$_{min-max}$ (↑)
LLM-agg	0.127±0.054	0.273±0.098	0.584±0.152	0.199	0.721±0.156	0.746±0.034	0.921
additive	0.099±0.035	0.238±0.073	0.623±0.11	0.075	0.640±0.071	0.720±0.023	0.914
least-misery	0.127±0.068	0.259±0.117	0.641±0.182	**0.480**	0.475±0.259	0.717±0.049	0.892
Borda	0.120±0.057	0.291±0.123	0.536±0.196	0.147	0.709±0.078	0.732±0.031	0.926
GreedyLM [12]	0.100±0.066	0.244±0.151	0.610±0.240	0.192	0.744±0.087	0.748±0.029	*0.932*
GFAR [2]	0.114±0.055	0.288±0.121	0.536±0.189	-0.083	0.761±0.083	0.751±0.029	*0.932*
EP-FuzzDA [3]	0.117±0.056	0.285±0.126	0.546±0.199	0.222	0.747±0.082	0.746±0.031	*0.932*
MCTS-diversity	0.105±0.053	0.263±0.118	0.576±0.189	0.239	**0.872**±0.027	0.779±0.022	**0.965**
MCTS-novelty	0.125±0.055	0.295±0.111	0.535±0.174	0.330	0.737±0.104	**0.806**±0.023	*0.936*
MCTS-MRR	0.108±0.054	0.269±0.122	0.568±0.196	0.253	*0.867*±0.031	*0.790*±0.023	**0.961**
MCTS-fairness	0.105±0.049	0.262±0.109	0.578±0.177	0.359	0.676±0.095	0.722±0.039	*0.931*
MCTS-weighted	0.120±0.057	0.287±0.119	0.546±0.189	0.292	0.742±0.112	**0.809**±0.023	*0.937*
MCTS-LLM	*0.140*±0.061	*0.317*±0.123	*0.508*±0.192	0.352	0.666±0.081	0.710±0.031	0.927
MCTS-LLM-rank	0.128±0.039	0.285±0.042	0.560±0.080	*0.440*	0.681±0.049	0.724±0.014	*0.939*
MCTS-LLM-pos	**0.148**±0.031	**0.344**±0.071	**0.460**±0.128	0.340	0.700±0.067	0.735±0.034	0.941

Table 1: Summarized mean group scores ($k = 5$) (best results in bold, second best in *italic*)

A cut-off threshold was set to select the top-K recommended items, with $K = 5$. Metrics were computed at the individual user level, and then aggregated to obtain the group score. For each user, recommendations were considered correct if they appeared in their test set. Group-level results were summarized from user results using metrics such as minimum, mean, median, variance, and maximum scores [2, 5, 6, 9]. To assess the statistical significance of differences, paired samples tests ($\alpha = 0.05$) were used, including corrections for multiple tests.

2 ANALYSIS

Table 1 presents the mean (and corresponding std) results for groups of 5 members, created using the similar group strategy. The metrics for each group were computed as the mean score for each group member. In the case of correlations, a meta-analysis [10, 11] was performed to compute their mean score. Inspired by [12] we also included the min-max ratio for novelty. In the case of precision and nDCG as the minimum was 0 in most cases, the resulting min-max ratio was also 0.

The MTCS was run for 200 iterations, and no simulation strategy was used. Due to the random nature of some MTCS strategies, we ran the evaluations multiple times using different seeds.

2.1 RQ1. Baselines versus MCTS

For different group sizes, as reported by Kaya et al. [2], the lowest recommendation relevance was observed for the largest groups, while the highest relevance was found in the smallest groups. The smallest differences among strategies were also seen for the smallest groups, with no alternative showing a distinctly higher performance. For groups of 4 and 8 members, the MCTS variants showed significant differences, with a few exceptions in precision and correlation. Similar trends were observed for divergent and random strategies for group creation with respect to the similar group strategy. Overall, scores were slightly higher for divergent and random groups, with the most significant differences seen in the divergent strategy.

2.2 Exploratory analysis of simulation strategies and number of iterations

We evaluated increasing the number of iterations from 200 to 500, 1000 and 1500. Although in general, increasing the number of iterations slightly modified the results, differences were not significant. The only exceptions were for *MCTS-diversity*, for which using 200 iterations achieved significantly better precision and nDCG (with small effect sizes), while increasing the number of iterations allowed to significantly increase the diversity and novelty of recommendations (with medium to large effect sizes). For divergent, similar tendencies were observed, with the exception that also *MCTS-novelty* saw its diversity/novelty increased.

We also evaluated introducing a random simulation strategy by which the last 2 elements in the recommendation sequence (of size 5) were simulated. In general, the observed differences were not statistically significant. There were only a few exceptions. For example, introducing simulations allowed MCTS-diversity to increase nDCG (with a small effect size), while it also caused a decrement of novelty (*MCTS-weighted* and *MCTS-MRR*) and precision (*MCTS-LLM-Pos*).

ETHICAL STATEMENT

While LLMs can enhance recommendation accuracy and user experience, they also raise concerns about fairness and bias. Ensuring transparency, accountability, and fairness in LLM-based recommendation systems is crucial for addressing ethical aspects and maintaining user trust.

REFERENCES

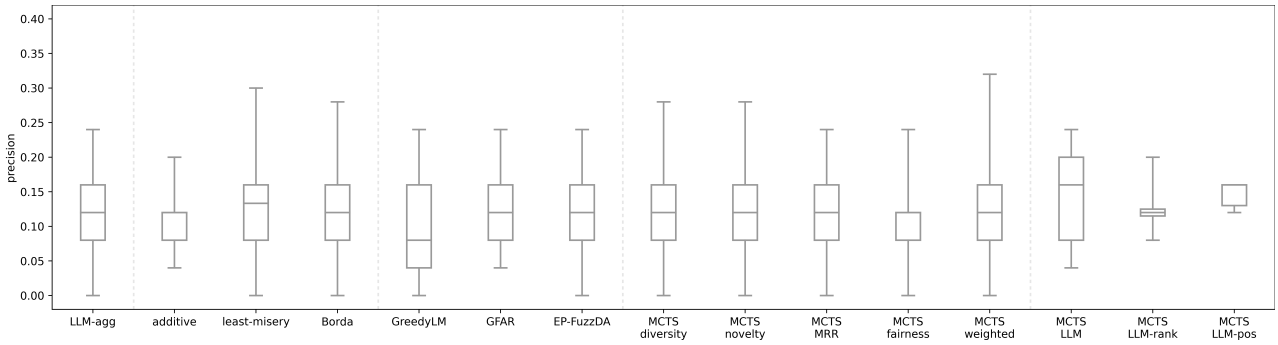
[1] Linas Baltrunas, Tadas Makcinskas, and Francesco Ricci. 2010. Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*. 119–126.

[2] Mesut Kaya, Derek Bridge, and Nava Tintarev. 2020. Ensuring fairness in group recommendations by rank-sensitive balancing of relevance. In *Proceedings of the 14th ACM Conference on recommender systems*. 101–110.

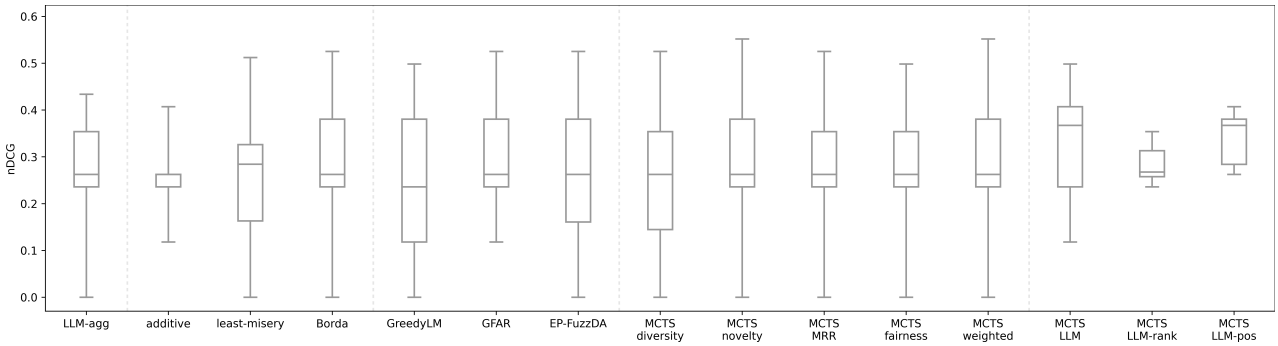
[3] Ladislav Malecek and Ladislav Peska. 2021. Fairness-preserving group recommendations with user weighting. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 4–9.

[4] Judith Masthoff. 2010. Group recommender systems: Combining individual models. In *Recommender systems handbook*. Springer, 677–702.

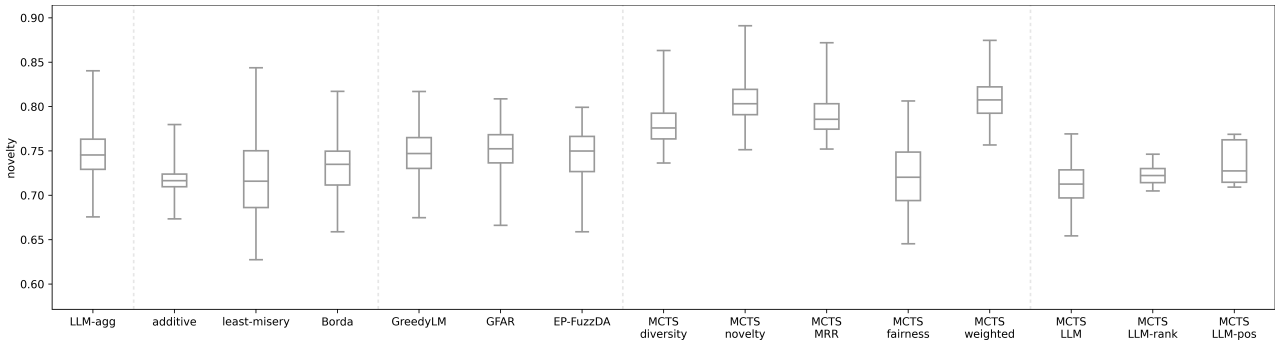
[5] Ladislav Peska and Ladislav Malecek. 2021. Coupled or Decoupled Evaluation for Group Recommendation Methods?. In *Perspectives@ RecSys*.



(a) Mean precision distribution across groups



(b) Mean nDCG distribution across groups



(c) Mean novelty distribution across groups

Figure 1: Mean metric distribution across groups

[6] Dimitris Sacharidis. 2019. Top-n group recommendations with fairness. In *Proceedings of the 34th ACM/SIGAPP symposium on applied computing*. 1663–1670.

[7] Javier Sanz-Cruzado and Pablo Castells. 2018. Enhancing Structural Diversity in Social Networks by Recommending Weak Ties. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. ACM, New York, NY, USA, 233–241. https://doi.org/10.1145/3240323.3240371

[8] Javier Sanz-Cruzado and Pablo Castells. 2018. Enhancing structural diversity in social networks by recommending weak ties. In *Proceedings of the 12th ACM conference on recommender systems*. 233–241.

[9] Christoph Trattner, Alan Said, Ludovico Boratto, and Alexander Felfernig. 2023. Evaluating group recommender systems. In *Group recommender systems: an introduction*. Springer, 63–75.

[10] Robbie CM van Aert. 2023. Meta-analyzing partial correlation coefficients using Fisher's z transformation. *Research Synthesis Methods* 14, 5 (2023), 768–773.

[11] David A Walker. 2003. JMASM9: converting Kendall's tau for correlational or meta-analytic analyses. *Journal of Modern Applied Statistical Methods* 2 (2003), 525–530.

[12] Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. 2017. Fairness-aware group recommendation with pareto-efficiency. In *Proceedings of the eleventh ACM conference on recommender systems*. 107–115.