RecSys 2024: Notification for Submission 513

1 message

RecSys 2024 <recsys2024@easychair.org> Mon, Jul 29, 2024 at 5:11 PM
To: Antonela Tommasel <antonela.tommasel@isistan.unicen.edu.ar>

Dear Antonela Tommasel,

Congratulations! We are happy to inform you that your SHORT paper ID: 513 - “Fairness Matters: A look at LLM-generated group recommendations” has been accepted for publication in the RecSys 2024 proceedings and for poster presentation at the conference.

The reviews are included below, which provide input for improving the paper for the camera-ready version. Formatting and logistics instructions for preparing your camera-ready version will be sent to you shortly in another email from our proceedings chairs. The deadline for submitting the camera-ready version is August 19, 2024.

RecSys 2024 requires at least one author to register and participate in the conference in-person. You will find more information about the registration process on the conference website (https://recsys.acm.org/recsys24/registration/). Papers for which no author has registered may be withdrawn from the conference proceedings.

RecSys 2024 is committed to providing attendees with an inclusive, safe, and respectful conference environment. We invite you to visit the RecSys 2024 Inclusion page (https://recsys.acm.org/recsys24/inclusion/) for more information on Diversity, Inclusion and Accessibility, including registration discounts and child care grants.

Thank you again for submitting to RecSys 2024. We look forward to seeing you and your co-authors in Bari in October!

Best regards,
Thorsten Joachims & Katrien Verbert
RecSys 2024 PC Chairs

SUBMISSION: 513
TITLE: Fairness Matters: A look at LLM-generated group recommendations

---------- METAREVIEW ----------
Overall we see more value than weakness in this paper.
The paper presents a study on a timely topic and has insightful observations, which could be interesting to the research community.
We suggest the authors to carefully take into account the review with negative rating and revised the final version properly.

---------- REVIEW 1 ----------
SUBMISSION: 513
TITLE: Fairness Matters: A look at LLM-generated group recommendations
AUTHORS: Antonela Tommasel

---------- Strong Points ----------
+ tackles a timely topic
+ code is shared to reproduce the results presented in the paper.
---------- Weak Points ----------
- some discussion is missing
- some parts needs to be more detailed.
---------- Overall Evaluation ----------
SCORE: 1 (weak accept)
----- TEXT:
Author(s) investigate LLMs impact on group recommendation fairness, establishing and instantiating a framework that encompasses group definition, sensitive attribute combinations, and evaluation methodology.

To the best of my knowledge authors cite relevant publications that studied fairness in RecSys and LLMs.

In Section 3.1 by referring to previous work author(s) also simulate groups of 2 to 8 members. I believe the effect of group size on the experimental results is very interesting aspect to look at. I am aware that this is a short paper but at least in the supplementary material or briefly in the paper this could have been discussed. At the end of the same Section, more discussion on why primarily, author(s) focus on similar groups instead of random or divergent groups is needed. Again, yes this is a short paper but a justification would be good. For instance, considering the fairness aspect of it, ideally 2 users having similar preferences should be satisfied equally no matter what their sensitive attribute is. Thinking from the group dynamics, the fairness issue would be more clearly visible for the random and divergent groups for instance. Some discussion on this would make the paper stronger.

In Section 3.3, for the Correlation of recommendations which Correlation metric did the author(s) use? Adding the name here would improve the readability.

One detail that I wonder for the Analysis Section is the reliability of the text generation. Which decoding strategies did the author(s) use for instance? Greedy search, beam search? This can also have an effect on the generated results. For instance, everything considered the same for the instruction whether LLM produces the same ranking or not.

Besides the concern listed above I believe this is a timely and interesting paper. Focusing on fairness of LLMs from the group recommendations is a novel research contribution.

---------- REVIEW 2 ----------
SUBMISSION: 513
TITLE: Fairness Matters: A look at LLM-generated group recommendations
AUTHORS: Antonela Tommasel

---------- Strong Points ----------
S1: The topic of this paper is relevant to the RecSys committee.

S2: The organization of this paper is clear.

S3: Experimental results validate the main claims of this paper.
---------- Weak Points ----------
W1: How to evaluate the correlation of recommendations? Is there any evaluation metric different from the rank correlations that can be used to evaluate the performance?

W2: Does the same phenomenon happen with other datasets (e.g. some short video datasets or e-commerce platform datasets)?

W3: Some typos: “Following prior research” should be “Following prior researches”, “As in previous works [1]” should be “As in previous work [1]”, etc.
---------- Overall Evaluation ----------
SCORE: 1 (weak accept)
----- TEXT:
Given the novelty and the clarity issue, I will give a "1" rating.

---------- REVIEW 3 ----------
SUBMISSION: 513
TITLE: Fairness Matters: A look at LLM-generated group recommendations
AUTHORS: Antonela Tommasel

---------- Strong Points ----------
1. This work studies on an interesting problem.
2. This work conduct empirical analyses to explore how LLMs affect the fairness of group recommendations.
3. The paper is well-written.
---------- Weak Points ----------
1. My major concern is on the problem definition. What is the formal definition of fairness in group recommendations? How does it differ from fairness in vanilla recommendation scenarios? What are the specific challenges? Providing a clear problem definition could help readers better understand the work and realize the necessity to study this problem.

2. Another concern is the experiments. The study only uses one dataset. Including more datasets would enhance the

generalizability of the conclusions.

3. This work misses some important work on LLM-based recommendation, including:

[a1] Large Language Models are Zero-Shot Rankers for Recommender Systems
[a2] Item-side Fairness of Large Language Model-based Recommendation System
---------- Overall Evaluation ----------
SCORE: 1 (weak accept)
----- TEXT:
This work investigates the issue of unfairness in LLM-based recommendations. While this work has certain limitations, it examines an interesting problem and provides some insightful conclusions. Consequently, my recommendation leans towards a weak accept.

---------- REVIEW 4 ----------
SUBMISSION: 513
TITLE: Fairness Matters: A look at LLM-generated group recommendations
AUTHORS: Antonela Tommasel

---------- Strong Points ----------
1) The paper investigated fairness related to sensitive attributes of group members in group recommendations, where the fairness is usually only focused on satisfaction of all the group members and not on sensitive attributes.
2) This is one of the first works proposing LLM-based Group Recommenders
3) The authors shared their code, enhancing reproducibility (but I suggest to check the link as you received a warning when opening it)
---------- Weak Points ----------
1) There are many aspects in the evaluation procedure that should be clarified
2) The work should also better be framed in the group recommendations SOTA, regarding fairness studies and privacy concerns, this would also highlight the novelty of focusing on sensitive attributes in this context
3) I think it would be important to discuss ethical considerations about this work: risks of providing sensitive attributes to the LLMs, privacy concerns (just mentioned at the moment), possible solutions.
---------- Overall Evaluation ----------
SCORE: -1 (weak reject)
----- TEXT:
The paper analyzes the impact of using LLMs for Group Recommendations, investigating how the fairness for the group is affected by biases that the LLMs could learn from the training data. I think this is a very interesting topic, and relevant for the Recsys community. I also think the work needs improvement, in most of the sections. This idea should be discussed in a long paper, in my opinion. It would be interesting to also investigate for specific differences depending on group size, group preferences similarity, distribution of sensitive attributes, for all the models, and not just reporting average results. In my opinion this work has a great potential and would really love to read more details. As it is now, I think there are many point to be improved, so I suggest to reject it.

Below more detailed feedback:

> Introduction
I think here Authors should make an effort to properly define the fairness aspects they investigate in the context of group recommendations. At the moment, it is not clear here if we consider fairness regarding all group members satisfaction, or fairness related to members belonging to "protected groups" (at the end, is more the second). The second aspect is under-investigated in group recommenders, and clarifying this would highlight the novelty of it. Furthermore, the contributions could be clarified in the end of the introduction.

> Related Works
I think this should be expanded. There are several works evaluating fairness for group recommenders, which usually focus on either the satisfaction of individuals in recommending sequences (see the already referenced work from Kaya et al. 2020 for an example) or evaluate fairness directly asking to real users (see Tran e al 2019 and Barile et al. 2024, but I also suggest reading Delic works on individuals satisfaction related to the group decision making process, will leave one as a reference). In this context, sensitive attributes are usually not considered, as the group recommendation is generated with aggregation strategies, the problem of sensitive attributes is more related to the individual recommenders that predicts individual ratings (I am simplifying for brevity). In the current work, however, the LLM is provided with individual and group "watching history", and sensitive attributes, hence these play a role. One could argue if it is actually necessary to provide the
LLM with this information, maybe Authors should add something about why doing this should improve the recommendations.

In the introduction, Authors mentioned privacy concerns in revealing sensitive data. There are also studies on privacy concerns regarding disclosing personal information in group explanations, which I think can be relevant for this point (I suggest to have a look to the works from Najafian for this, again I leave one at the end as reference) and this point could also be mentioned in the related works.
The last part of the related focusing on the LLMs for RS, explaining prompts used in related works, could be directly connected to the current work.

> Task and Method
Authors split the Movielens dataset 80/20 and transformed the ratings in binary labels (liked/disliked, if rating>=4). The generated groups are from 2 to 8, all with similar watching history. It is not explained why the focus is only on similar groups, which is the easiest case. It would be interesting to also evaluate the mentioned divergent or random groups.
Here I also have some explicit questions:
- Is the similarity evaluated only using movies in the training set? This should be specified.
- How the sensitive attribute "race" is associated to the users in Movielens? After several reads it seems this is missing in the dataset, and from the results it seems to me that both gender ad race are associated to the same groups in a controlled way (so the gender in the dataset is ignored), is this true?

Regarding the generation of the recommendation, I think a clarification is necessary about the provided set of movies to recommend, which includes: (i) all movies in the test set liked by all the group members; (ii) at least 5 exclusively liked by each member - what does it mean? 5 movies per group member they like and all the others dislike? there can be more than 5?; and (iii) movies disliked by each member - are these movies disliked by all group members? how many are included?

In the "neutral scenario" the user sensitive attribute are just not provided to the LLM?

In (3.3), it should be clarified what Authors consider "rank correlation": which correlation measure is used, how you correlates rankings (is this considering a correlation between scores used to rank the movies in a group test set? or the specific rank of each movie is used?).

> Analysis
It is very quickly mentioned that ImplicitMF is used in combination with Group preference aggregation. The reason is never introduced, this should be explained here. Also, which aggregation is used? How the ImplicitMF is trained? Is this specified somewhere?

RQ1: is there any info about the significance of these correlations? However, it is not surprising to me that providing more context to the LLM the resulting recommendations are different.

RQ2: In RQ2 Authors analyze precision maximum variance. The RQ2 mentions "individual preferences" but Authors actually analyze how recommendation quality varies. I think the formulation of RQ2 should be different and refer to precision variance or quality of recommendations, and not to individual preferences. ImplicitMF results are mentioned but these are not in any figure or table, neither in the text, are these in additional material? What does these results add?

RQ3: Here by "scores" you mean both precision and NDCG?

> Conclusions

I understand that the limited space is the motivation of this, but I think the ethical statement should be in the main paper, and more ethical considerations should be discussed.

REF (suggested readings)
- Tran et al. (2019). Towards social choice-based explanations in group recommender systems. UMAP.
- Barile et al. (2024). Evaluating explainable social choice-based aggregation strategies for group recommendation. UMUAI.
- Delic et al. (2018). An observational user study for group recommender systems in the tourism domain. IT&T.
- Najafian et al. (2023). How do people make decisions in disclosing personal information in tourism group recommendations in competitive versus cooperative conditions?. UMUAI.