Evaluating Biases in Group Recommender Systems: Challenges and Solutions.

Gabriel Feltes dos Santos[a].

[a] Computing Department, Science, Technology and Health Department, Federal University of Santa Catarina, Araranguá, Brazil, gabrielfeltes.gg@gmail.com.

Abstract. Group Recommender Systems (GRS) are a subdomain of classical single-user Recommender Systems (RS). They are applied in situations where a group of people need to combine their preferences in order to receive recommendations that aim to preserve both personal and overall satisfaction. This approach demands certain complexity regarding social interactions, fairness, divergent opinions and lack of sufficient data. GRS aggregators operate on top of RS infrastructure, because both of them have the task of generating single outputs or an ordered list of outputs. So far, the article demonstrates how to define a fairness parameter for evaluating the quality of recommendations for the group, in addition to how to personalize the weight of each user of the GRS according to their importance. Going further, there is a presentation of how to reduce the effects of biases generated from different parts of the GRS. Some results of these evaluations are presented based on the work of cited researchers in order to address challenges and possible improvements needed for the fair use of GRS by society.

Keywords. Group Recommender Systems, bias, evaluation, aggregator, fairness.

1. Introduction

The rise of digital platforms and the growing accessibility of online services have transformed the way people interact with information and each other. As a result, recommendation systems have become essential tools for customizing user experiences and helping individuals navigate vast amounts of content. Among these systems, Group Recommender Systems (GRS) play a critical role in providing tailored recommendations for groups rather than individuals.

Group Recommender Systems gather preferences from a cluster of users to generate sequential recommendations for the entire group. This approach aims to satisfy the collective preferences of the group while balancing individual tastes. However, developing effective GRS involves navigating a complex set of challenges, particularly concerning biases that can influence recommendations.

Two primary categories of GRS evaluation — coupled [1] and decoupled [2] systems — present different challenges when it comes to managing biases. In coupled GRS, popularity biases can lead to skewed recommendations that favor mainstream or trending content over less popular yet potentially relevant options. Meanwhile, decoupled GRS face polarity biases, where recommendations might lean heavily toward either positive or negative feedback, rather than providing a balanced perspective .

This article examines the origins of biases in Group Recommender Systems (GRS) and how they can spread across various system components. By understanding these challenges, we may be able to create more effective and equitable GRS that cater to the diverse preferences of different groups. This could improve user satisfaction and experience across a variety of applications.

2. Methodology

The crucial method for defining a state-of-the-art approach in terms of biases in GRS is a personalized literature review. Hence, let's navigate through the components of a GRS, from the elicitation of user preferences until the aggregation of multiple preferences into an ordered list.

2.1 Collecting data from users

According to [3], datasets can be built using three methods: offline, online, and user studies. The diversity of the dataset contributes to the conclusiveness of general hypotheses regarding algorithm efficiency. However, this efficiency is not static; it depends on contextual factors and variables considered during algorithm execution.

The offline method abstracts the expected behavior of recommender systems by employing atemporal and general definitions for their qualities. This allows newly proposed algorithms to undergo initial evaluation through offline filtering. Subsequently, the filtering process progresses to consider the contextual nuances through online and user study evaluations.

User studies aim to inform the public about the utilization of their data in algorithm training processes. On the other hand, online evaluations delve into mining user data, which is recorded in response to specific and detailed interactions via the human computer interfaces.

The issue here is that online evaluations are quite expensive for being conducted in academia [4]. Thus, some frameworks for user studies were developed to facilitate personalized and extensible tools to build datasets [4, 5]. These frameworks can be used to analyze biases and fairness within recommendations.

2.2 Avoiding presentation biases

When comparing the performance of algorithms in user studies, a list of recommended items to be selected by the user can have advantages over another one due to their relative position on screen [6]. This can lead to an unfair chance of items being selected by the user, which constitutes a presentation bias.

2.3 Combining user preferences

After managing the definition of datasets, we perform how user profiles will achieve consensus. Thus, GRS aggregators are used on top of single-user recommenders and can form either a group profile with its own preferences or an aggregation of the items previously recommended for each user. By considering the last method, we can define consensus and fairness for GRS [7].

Thus, if we consider some items that were rated in a 0.0 to 5.0 scale by all the users of a group, a measure of fairness can be inferred. As Fig. 1 shows, a rating threshold (or a minimum value for considering the rating a positive one) is defined as 3.5. For three items, the amount of ratings greater than 3.5 for each user is counted. If the amount is greater or equal to 2 (m-proportionality), the user is considered satisfied with the items presented to the whole group.

	Item ratings				
g_1	t_1	t_2	t_3	Rating threshold > 3.5	m-prop
u_1	5.0	4.0	1.0	2	1
u_2	4.0	3.0	2.0	1	0
u_3	4.5	5.0	5.0	3	1
u_4	3.5	3.0	5.0	1	0
$fairness_{m\text{-}prop}$	–	–	–	–	$\frac{2}{4} = 0.5$

Fig. 1 - Evaluating fairness based on m-proportionality, where m = 2.

2.4 Beyond fairness criteria

A simple definition of fairness may not adequately address the preferences of minorities. Balancing or attributing weights to each group member is essential for maintaining long-term fairness in groups with varying user importance. In relation to public elections, there are risks when using group recommendation systems (GRS).

The focus is on D'Hondt's algorithm and its extensions, which are greedy selection methods that choose the best candidates based on the volume of applicable votes. Using D'Hondt's algorithm directly in GRS can be challenging because it doesn't account for the differing relevance of items for different users or the possibility of item overlap across users' preferences [8].

To address these challenges, researchers have explored FuzzDA, a fuzzy extension of D'Hondt's algorithm that aggregates multiple recommendation systems proportionally. FuzzDA selects candidates that maximize the sum of accountable votes for all users and adjusts votes according to relevance scores given by each user.

EP-FuzzDA [8], an improvement on FuzzDA, addresses group recommendation challenges by considering both relevance and fairness. It introduces the Exactly-Proportional relevance sum (EP-rel-sum) criterion, which balances each user's proportional representation with the total relevance of items.

By limiting user-specific relevance, EP-FuzzDA ensures each user's influence on recommendations matches their share of total relevance. This approach helps maintain both relevance and fairness in recommendations, even with varying user preferences and overlapping interests. Evaluations show EP-FuzzDA's effectiveness in achieving fairness while optimizing total item relevance.

2.5 Mitigating biases from GRS ground truths

The evaluation of GRS aggregators can be divided into coupled [1] and decoupled [2] approaches. The first one considers the underlying RS (single user RS) and the aggregator as a tightly coupled pair, because the ground truth of the aggregator is the test set of the underlying RS. This leads to dependence on the performance of recommendations for single users, resulting in popularity bias perpetuation (popular items have more ratings and, thus, chances to be picked up). To mitigate this, the self-normalized inverse propensity score (SNIPS) is used to normalize the popularity bias present in a set of user's relevant items.

$$r^{SNIPS}_{L_G, u} = \frac{1}{\sum_{i \in R_u} \frac{1}{P_{u,i}}} \sum_{i \in R_u} \frac{score(L_G, i)}{P_{u,i}} \quad (1)$$

Equation 1 shows that the relevance of the list of recommendations (L_G) for the user u is estimated for

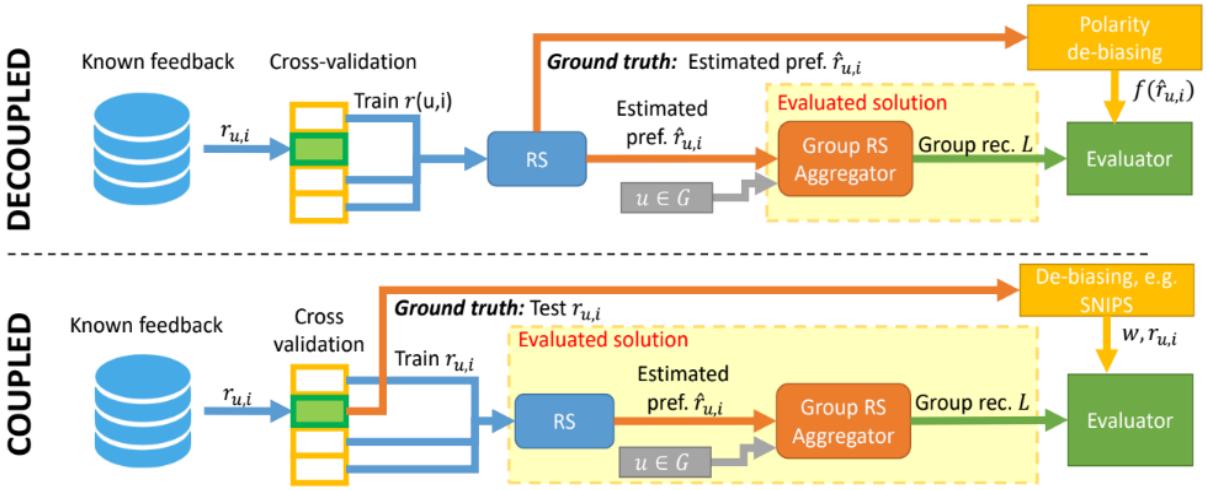

Fig. 2 - Coupled and decoupled schemas [2].

the set of user's known relevant items (R_u), estimated item's propensity score ($P_{u,i}$) and some scoring metric score (L_G, i). For further explanations, check [1].

Meanwhile, the decoupled approach uses the relevance score estimated by underlying RS as a ground truth for the evaluation of group aggregators [2]. This results in an overestimation or underestimation of the user's real preferences. To counteract this, the estimated preferences for each user are refactored with a polarity de-biasing technique. Check the differences between decoupled and coupled evaluation in Fig. 2 [2].

2.6 Behavioral sources of biases

More in-depth concepts regarding group dynamics influence the GRS priorities for designing recommendations. For instance, Felfernig et al. [9] presents the concept of anchoring: when the decisions of some group members are disclosed to other ones who are still thinking about their preferences, these undecided ones tend to be influenced or pressed to make a decision. Such behavior was observed in an important context of software requirements engineering. The earlier user-individual preferences are shown to other users, the less ratings of users differed, in terms of standard deviation [9]. On the other hand, when the disclosure of preferences happens after a longer time period, the discussion among users intensifies and has a positive impact on decision quality.

Following the importance of fostering information exchange in group decision processes, there are some groups that avoid important conflicts. These ones should be integrated by experts outside the group in order to stimulate diverse opinions. Again, evidence shows that not allowing personal contagion at early stages of group decisions is crucial [9].

Another very important factor to be considered is the effect of polarization in group decisions. Individual ratings and opinions tend to become more risky or polarized when the group encourages individuals to take their side. It should be counteracted by triggering discussions related to the negative impacts of these biased and possibly dangerous decisions. Political parties around the world tend to be polarized because of the concern of having an unique and opposite side to be defeated. Group Recommender Systems offer insights about this great issue.

3. Results

Both coupled [1] and decoupled [2] evaluations had their results well detailed in their papers. Their graphics are a result of Python libraries that plot results of evaluations after all the database population and algorithm training. The hyperlinks for the repositories can be found in both works, as well as the main graphics for analyzing results.

Basically, some algorithms were filtered in an offline evaluation with the aim to identify which ones provide best GRS performance for either coupled or decoupled methods. The use of EP-FuzzDA [8] served to compare its performance with the ones from other previous algorithms in terms of uniform weighting scenario, weighted scenario and long-term fairness. By combining these metrics, the algorithms can be ranked for each category and context of GRS evaluation.

4. Discussion

In the methodology section, the methods of evaluating GRS were depicted into a filtering process for which some algorithms were ranked in order to prove the efficiency of EP-FuzzDA [8]. However, this algorithm was only evaluated for static datasets under offline approach. If we consider the case of presentation bias [6] related to a collection of user data from user study frameworks [4, 5], there will be more issues to be discussed about fairness.

As stated in section 2.6, there is a crucial factor for maintaining the authenticity of users in a GRS: the privacy of their choices in the early stages of the

recommending session. The actions we take are highly influenceable by others' already taken decisions. Hence, the user interface unveils its potential to manage power relations among users.

5. Conclusions

In conclusion, understanding and addressing biases in Group Recommender Systems (GRS) is essential for creating fair and effective recommendations. Biases can arise from how data is collected, how recommendations are presented, and how user preferences are aggregated. Both coupled and decoupled evaluation methods offer different approaches to handle these challenges, and newer methods like EP-FuzzDA present promising solutions.

To achieve fairness, it is important to balance individual and group preferences while managing issues like polarization and anchoring within groups. Considering human behavior and group dynamics is crucial in designing recommendation algorithms that cater to diverse user needs.

By developing strong methods to evaluate and mitigate biases, we can improve user satisfaction and experiences in GRS. As the field advances, ongoing research and innovation will help navigate the complexities of group decision-making and deliver more personalized and meaningful experiences across various digital applications.

6. References

[1] Peska, L., & Malecek, L. Coupled or Decoupled Evaluation for Group Recommendation Methods? In: Zangerle, E., Bauer, C., Said, A., editors. Proceedings of the Perspectives on the Evaluation of Recommender Systems Workshop 2021 co-located with the 15th ACM Conference on Recommender Systems (RecSys 2021), Amsterdam, The Netherlands, September 25, 2021. CEUR Workshop Proceedings, Vol. 2955. http://ceur-ws.org/Vol-2955/paper1.pdf

[2] Dokoupil, P., Peska, L. Robustness Against Polarity Bias in Decoupled Group Recommendations Evaluation. Proceedings of the Perspectives on the Evaluation of Recommender Systems Workshop 2022 co-located with the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22 Adjunct), Barcelona, Spain, July 11, 2022. CEUR Workshop Proceedings, Vol. 3288. https://doi.org/10.1145/3511047.3537650

[3] Gunawardana, A., Shani, G., Yogev, S. (2022). Evaluating Recommender Systems. In: Ricci, F., Rokach, L., Shapira, B. (eds) Recommender Systems Handbook. Springer, New York, NY. https://doi.org/10.1007/978-1-0716-2197-4_15

[4] Dokoupil, P., & Peska, L. (2023). EasyStudy: Framework for easy deployment of user studies on recommender systems. Proceedings of the 17th ACM Conference on Recommender Systems, Singapore, Singapore. Association for Computing Machinery, New York, NY, USA, 1196–1199. https://doi.org/10.1145/3604915.3610640

[5] Anelli, V. W., Bellogin, A., Ferrara, A., Malitesta, D., Merra, F. A., Pomo, C., Donini, F. M., & Di Noia, T. (2021). Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada. Association for Computing Machinery, New York, NY, USA, 2405–2414. https://doi.org/10.1145/3404835.3463245

[6] Dokoupil, P., Peska, L., & Boratto, L. (2023). Rows or Columns? Minimizing Presentation Bias When Comparing Multiple Recommender Systems. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei, Taiwan. Association for Computing Machinery, New York, NY, USA, 2354–2358. https://doi.org/10.1145/3539618.3592056

[7] Trattner, C., Said, A., Boratto, L., Felfernig, A. (2024). Evaluating Group Recommender Systems. In: Felfernig, A., Boratto, L., Stettinger, M., Tkalčič, M. (eds) Group Recommender Systems. Signals and Communication Technology. Springer, Cham. https://doi.org/10.1007/978-3-031-44943-7_3

[8] Malecek, L., & Peska, L. (2021). Fairness-preserving group recommendations with user weighting. Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, Utrecht, Netherlands. Association for Computing Machinery, New York, NY, USA, 4–9. https://doi.org/10.1145/3450614.3461679

[9] Felfernig, A., Atas, M., Stettinger, M., Tran, T.N.T., Leitner, G. (2024). Biases in Group Decisions. In: Felfernig, A., Boratto, L., Stettinger, M., Tkalčič, M. (eds) Group Recommender Systems. Signals and Communication Technology. Springer, Cham. https://doi.org/10.1007/978-3-031-44943-7_8