Communication Papers of the of the 17th Conference on Computer Science and Intelligence Systems pp. 47–54
DOI: 10.15439/2022F75
ISSN 2300-5963 ACSIS, Vol. 32

Evaluating Diversification in Group Recommender Systems*

Amanda Chagas de Oliveira
Institute of Computing
Federal University of Bahia

Frederico Araujo Durao
Institute of Computing
Federal University of Bahia

***Abstract*—The formation of groups is an ordinary event in our routines. For example, people used to lunch, travel, or hang out in groups. Conversely, getting a consensus over an item may be difficult for some groups as the number of digital information increases. Group Recommender Systems (GRS) rise to assist in this task, as they filter which items may be more relevant to the group. Although there are consensus techniques to help in this matter, recommendations to groups can become monotonous, and this opens space for applying diversification techniques to improve recommendations. In this paper, we expose a model for recommendation to groups using diversification techniques and present the results of the online experiment where the proposal obtained an increase in precision at all levels compared with baseline.**

I. Introduction

RECOMMENDER Systems (RS) are automated tools for locating the information that is pertinent to the user [1]. Although most RS are made to serve a single user, there are instances where a group of interests must be considered. Therefore a Group Recommender System (GRS) must consider each member's preferences.

A GRS has the role of finding what is relevant to a group rather than to individuals. Given that we live in communities participating in group activities is a usual behavior in everyday life. However, even straightforward tasks such as selecting a playlist for a group of friends can be challenging. All individual preferences must be considered when processing recommendations in a group scenario. The group size at least multiplies the recommendation problem. Therefore, it is necessary to use consensus techniques to identify the items that satisfy the group as a whole as well as each member. In the literature, GRS has been examined from a variety of angles. [1] investigate strategies for aggregating individual preferences, [2] study on effective group recommendation methods, and [3] use various strategies when recommending items to groups.

Group recommendations are based on a group profile that combines the members' preferences. Recommendations frequently fall into known scenarios without diversity if such a profile is not kept updated or modified. This issue, known as overspecialization, may negatively affect GRS since recommendations become increasingly repetitious and unappealing. Hence, GRS must exploit various approaches to diversify recommendations to improve the overall group satisfaction. Diversity in RS has been extensively researched in the literature in recent years. The first work to formally introduce the concept of diversity was [4], [5] explore diversity while evaluating RS, and [6] discuss the effects of diversity algorithms in group recommendations. By addressing the issue of overspecialization in diversity for group recommendation, this article seeks to contribute to this particular field of research.

Therefore, in addition to boosting diversity in the recommendation list, we suggest creating and evaluating a group recommendation model that employs a diversification algorithm to optimize consensus among members. The main objective therefore is to lessen the effects of overspecialization to maximize group members' satisfaction. This proposal is an extension of a previous work [7], on which the authors developed the preliminary version of the group recommendation model using diversification techniques. Unlike the previous work, this proposal provides a comprehensive user assessment that addresses crucial aspects involving group recommendations such as group size and group formation, besides discussing the results from real users. In particular, we perform a user trial in which 6 groups of users assess the recommendations generated by our model against a state-of-the-art baseline method. Because the groups differ in size, we also discuss how this attribute impacts the precision of recommendations.

The research questions that drive our study are i) Are the group recommendations still relevant even after diversification? and ii) How accurate are the predictions considering the error?

This paper is structured as follows. Section II presents related work. Section III provides important background on the area, Section IV depicts the overall approach. Section V presents the experimental evaluation. Section VI discusses the key achievements and points out the limitations of the work. Section VII concludes the paper and sets forth the future works.

II. Related Work

GRS generally suggests a group of individuals participating in a group activity. Group recommendation has received much study in the literature from a variety of angles, including aggregation techniques [1], [5], group consensus [3] and graph-based algorithms [8]. However, the use of diversity algorithms in GRS is yet a low explored field [6].

The authors would like to thank FAPESB and CAPES for the financial support. Grant Term: PPF0001/2021. Technical Cooperation Agreement 45/2021. CAPES Program: PDPG-FAP no. 88887.637752/2021-00.

Regarding GRS, [1] provide a thorough review and insightful explanations of a number of aggregation techniques, including *Least Misery*, *Most Pleasure* and *Average Without Misery*. [2] propose semantics that accounts for item relevance and group disagreements. Nonetheless, use three group building strategies to bring users together: first similar users, then dissimilar users, and finally, groups formed randomly. In our work, the users create groups as they wish, as detailed in Section IV-B. [9] propose to utilize the multi-criteria ratings to learn the group expectations to build their effective group recommendation models. The author presents a new dataset related to the educational field where group preferences are already collected, thus dispensing the use of group formation strategies. [8] focus on generating recommendations to massive groups, varying from 10 to 1000 members. The authors explore the group interest and the connection between group users to divide a big group into subgroups and generate an interest subgroup-based recommendation list. Thus, the generated lists are aggregated into the final one by a dynamic aggregation function considering the subgroup's contribution. In this work, we do not form groups based on their connections or similarity, and our experiments were conducted with real users rather than synthetic ones. [10] present a method for group recommendation in Telegram. Their method receives a set of users, analyses their groups, and recommends a list of ranked groups. Considering the membership graph and users record, the authors combine two previous pieces of research to achieve their proposed method. In contrast to this work, the author's proposal does not recommend items to the groups. They rather generate ranked groups using graph-based algorithms.

Diversity in RS is discussed by [5], which indicates significant concerns beyond RS accuracy. They highlight the advantages of diversification algorithms in particular and give numerous methods for re-ranking recommended lists. Similarly, we chose this course of action from a group standpoint. Diversity is a broad concept. The authors in [11] present a latent factor model that achieves the required accuracy level while maintaining a certain level of diversity in RS. The authors use elastic-net regression to regularize the model for accuracy and diversity in an optimization framework. However, in this work, we address the diversity topic from a group perspective and perform a greedy re-ranking in the final recommendation list. In addition, [12] suggest a method to re-rank the recommendation list by appearance frequency of items to recommend more range of items to improve diversity. Their appearance frequency score is calculated over the user's rating predictions. [6] also discuss the problem of diversity in GRS. However, their experiments ran over synthetic groups rather than real ones. However, it is also not clear how their preference matrix is set up.

III. Background

A. Aggregation Techniques

Aggregation techniques in a GRS are consensus functions capable of integrating different preferences into a single one or the group profile. The proposed method combines individual scores to create a collective profile. The aggregation techniques used in this proposal, the Average Without Misery (AWM), was inspired by [1].

The approaches of aggregation are [3]:

- **Average Without Misery (AWM).** This technique can be characterized as a synthesis of the Average and Least Misery strategies (LM). The Average determines the mean of each individual's ratings for each candidate item, taking the average into account as the group rating for that particular candidate item. The LM approach assumes that the group rating for that candidate item is the lowest individual rating, hence sparing group members from misery. The AWM was implemented in this study as follows: LM specifies the group rating for each candidate item if any individual rating for that candidate item is equal to or less than the threshold. However, the Average establishes the group rating for that candidate item if the individual ratings are above the threshold. We define the threshold as two based on the Likert Scale [13] when considering the dataset used in this study, in which ratings range from 0.5 to 5. In this instance, value 2 already conveys disapproval.

Most studies in the literature use the Average and Least Misery techniques. In this paper, we implement the AWM as a combination of both. From the results obtained in our previous work [7], the AWM technique performs better than the others analyzed.

B. Diversification Algorithms

Diversification can be defined for a list of items as a factor expressing how different pair items are on this list [4].

$$Similarity(x, y) = \frac{\sum_{i=1}^{n} w_i \cdot sim_i(x_i, y_i)}{\sum_{i=1}^{n} w_i} \tag{1}$$

Equation 1 defines the similarity between a pair of items, where n is the item attributes, w is the weight of the attribute, and $sim(x, y)$ is the comparison of attribute i from items x and y. The **title** and **genres** are attributes from the items that are used in the Cosine Similarity similarity calculus in this paper.

In particular, this proposal assessed two ways to diversification from the literature: Bounded Random Selection and Bounded Greedy Selection.

- **Bounded Random Selection.** On this algorithm, there is a list L with the rated items by the user, a list of candidate items C, and the final list with diversified recommendation R. For each item i_l in L, the algorithm searches for items in C similar to i_l and adds those items in a new list J, with a bounded length. Then, items are randomly chosen in J and included in R.
- **Bounded Greedy Selection.** This approach selects items greedily by taking the most diverse item on each turn and appending it to R. However, it is essential to define a greedy selection function. This greedy function needs

to consider similarity and diversity of the items. The flow of this algorithm is similar to the Random; however, the greedy function selects the items that maximize diversity while still taking similarity into account, rather than picking them at random from J.

$$Diversifying(x, R) = \alpha \cdot rel(x) + ... \\ ... + (1 - \alpha) \cdot \frac{1}{|R|} \sum_{y \in R} dist(x, y) \quad (2)$$

Equation 2 in this study presents the greedy function for weighting diversity and similarity [5], where α is used to balance the equation's factors, and $rel(x)$ is the relevance function for item x, which can be expressed on similarity. This equation represents the greedy approach for diversification covered in [14], [15], and [5].

IV. A Group Recommendation Model Using Diversification Techniques

A. Notations and Proposal Flow

The proposed model recommends a list of movies I to a group G composed of n users $u \in U$.

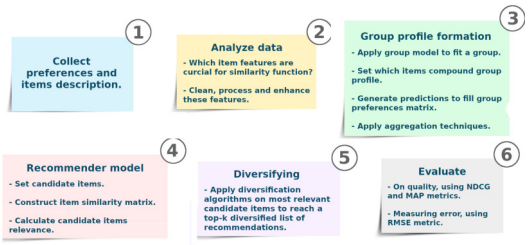

Fig. 1. Flow of proposed model defined in steps.

Figure 1 enumerates the recommendation flow. The 1st step is collecting item descriptions and user preferences from a given dataset. The 2nd step is about data prepossessing to reduce noise and redundancy. In the 3rd step, the group profile is created using aggregation techniques. In the 4th step, the group recommendations are generated. Diversification takes place in step 5. Finally, the group recommendations are evaluated in step 6. In the following, steps 3, 4, and 5 are depicted.

B. The Group Model

A group can be defined as a system of recurrent social relations or a reunion of people who share some characteristic, some idea, or some common interest [16]. Hence, it is crucial to define rules for group formation when recommending to groups. We divided the rules into two: group size and cohesion between members. **Group size:** The size of the group can vary drastically even in real life, from a couple having lunch to a crowd of thousands in a football stadium. In our experiments, we set the group size to 3 or 5, also inspired by recurrent use in the literature [17] and [6]. **Cohesion between members:** This aspect focuses on the relationship between members. Most of the datasets used in RS focus on modeling individual preferences rather than the relationship between those users. In our experiments, we asked participants to form groups as they wished.

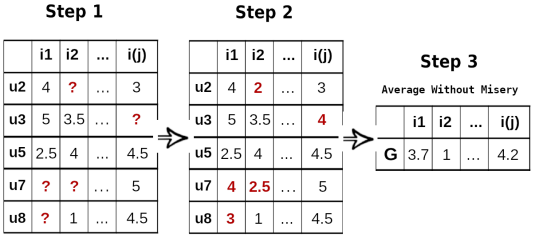

Fig. 2. Group profile generation flow.

Once the group is created, the next step is to define the group profile. In this phase, aggregation techniques are applied to individual preferences toward a single group profile G.

Figure 2 illustrates 5 members u_2, u_3, u_5, u_7, and u_8, and their respective ratings assigned to each item i_1 to $i_{(j)}$. The symbol ***?*** indicates no rating.

In order to apply aggregation techniques on the user preference matrix, the vacant slots (***?***) are predicted. We tested different prediction algorithms: Neighbourhood-based [18] i) *KnnWithMeans*, ii) *Knn*; and Matrix Factorization Based iii) Singular Value Decomposition (SVD) [19]. From our empirical observations, we selected the SVD algorithm which performed more accurately. Figure 2 - Step 2 illustrates the dense user preference matrix after employing the SVD algorithm. Once the user preference matrix is dense, the aggregation techniques are applied. Figure 2 illustrates the aggregation technique used in this paper $AverageWithoutMisery$, thus producing the group profile $G = \{3.7, 1, ..., 4.2\}$.

C. Recommendation Model

To generate group recommendations, we compare each movie's title and genres against candidate items. We only used these two features because information like director or cast were not available in the dataset. Therefore we employ a Content-Based (CB) approach [20] that recommends unknown items that are similar to the better-rated items in the group profile (see Algorithm 1).

In Algorithm 1, pi and ci stands for profile items and candidate items respectively. GP is the group profile and it is sorted by ratings, bc refers to the best candidates and stores the most similar candidate items from profile items. The list R is the recommendations generated for the group, and DR is the final list with the recommendations diversified. Important structures in Algorithm 1 are detailed as follows:

.

Algorithm 1: Algorithm for Group Recommendation.

```
Data: n as the number of groups; size; Users as all dataset
      users; Items as items from dataset; sim as the
      similarity matrix of items.
Result: DR list of recommendations diversified.
1  while n > 0 do
2      G = random(Users, size)
3      pi, ci = splitDatasetItems(Items, G)
4      GP = generateGroupProfile(pi, G)
5      bc = getMostSimilarItems(GP, ci, sim)
6      for x ∈ bc do
7          for y ∈ GP do
8              r = relevance(x, y)
9              R = R + r
10         end
11     end
12     DR = diversifyRecommendations(R)
13     n = n - 1
14 end
```

1) **Set candidate items.** It is important to split the dataset out into candidate and profile items. As mentioned in Section IV-B, any item that is rated by a group member is classified as a profile item, otherwise, it is a candidate item. This step is viewed in line 3, where pi are the profile items and ci are the candidate items.
2) **Sorting the group profile.** Provided the group profile, we know which items the group enjoys at most. We sort this list in a descending order by ratings $r_{G,i}$ in order to keep the preferred items at the beginning of the list. This step is defined in line 4.
3) **Constructing the items similarity matrix.** The next step is to build up the similarity matrix between items, where each cell value (x, y) corresponds to the similarity score for items x and y. Particularly, we use Cosine Similarity [21] as it performs great with textual information. The experiment for weighing the cosine similarity was based on the returned items' relevance, i.e., which movies are more similar to the group profile. We performed several tests empirically until achieving the best relevance. Then we applied the following setting: 0.8 for the title and 0.2 for the genre. Algorithm 1 receives the similarity matrix as an input sim.
4) **Setting relevance.** At this stage of the model, we already have the group score over the profile items, and we know which candidate items are more similar to profile items bc. Therefore, it is crucial to quantify how relevant a candidate item is to a group. Thus, we elaborate the Equation 3 that combines similarity values with group preference to express the relevance of the item.

$$relevance(x, y) = \frac{a \cdot sim(x, y) + b \cdot \frac{r_{(G,y)}}{max(r)}}{a + b} \quad (3)$$

Equation 3 shows x as a candidate item from bc, and y as a preferred item from GP for group G. In the first factor of the equation, the similarity between items, $sim(x, y)$ is normalized as well as the second factor $r_{(G,y)}$. Then, the relevance of x is the result of the weighted mean of variables a and b over x and y similarity, and how preferred y was rated by G, respectively. The sum of variables a and b may never be less than 1. This step is defined in line 8 of the algorithm.

D. Diversifying Group Recommendations

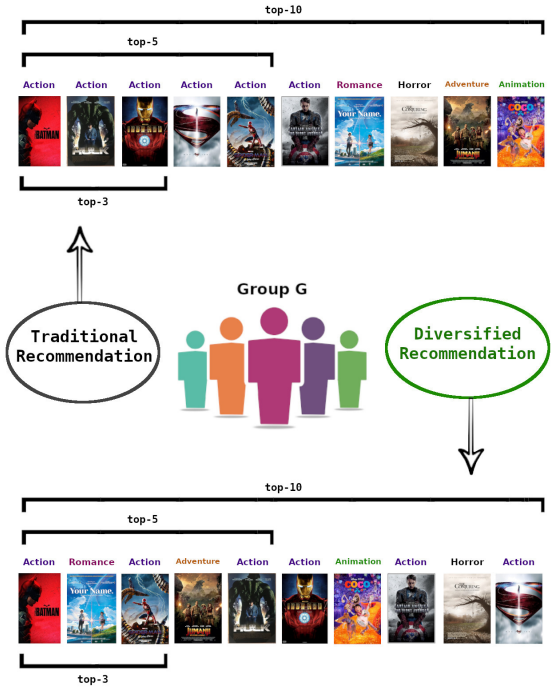

Fig. 3. Impact of diversification on top-k.

The outcome of the recommendation model is a list of recommended items R for the group. However, how diverse is R? In other words, how similar the top-k items are on this list? Considering k as 10, most items tend to have similar aspects. Therefore, we diversify the top-k to generate a relevant and diverse ranking for the group.

Most of the diversification algorithms are based on re-ranking items to improve diversity. Likewise, we implement the *Greedy Re-ranking Algorithm* [5]. This approach is used in Algorithm 1 at line 12. The entire implementation of this proposal is available at an online repository [1].

Figure 3 illustrates the recommendations for group G, using traditional recommendation techniques versus diversified recommendations. By observing the traditional recommendation set, up to the top-5 items, all movies belong to the same genre, i.e., action. A different genre is only observed in the item at the 7th position. In contrast, some diversity is already observed in the top-3 list using the diversification approach. It is important to outline that both lists are formed by the same items, the second an output from the diversification process.

V. Online Experimental Evaluation

In order to evaluate our model with real users, an online experiment was undertaken. This section depicts the method-

[1] https://github.com/amandachagas/GRSwithDiversity.git

TABLE I
DATA SAMPLE USED IN THE EXPERIMENT.

Id	Title	Genres	Year	ImdbLink	Cover	Youtube
1	Daybreakers	Action, Drama, Horror...	2010	...title/tt433362	...images-na.ssl...	.../watch?v=CtiLjvVwvY4
2	Jupiter Ascending	Action, Adventure, Sci-Fi	2015	...title/tt1617661	...images/MV5B...	.../watch?v=THVFkk-sEus
3	The Commuter	Crime, Drama, Mystery...	2018	...title/tt1590193	...images/Adh4F...	.../watch?v=aDshY43Ol2U

ology, dataset, metrics, and results obtained. In the end, we discuss the results and the overall achievements.

A. Methodology

The online experiment was conducted with 24 student volunteers at the University. Before participating in the experiment, they were introduced to the GRS concepts and the experiment's goals. We also ensured that all personal information was anonymized and used strictly for scientific purposes. All were informed about the two phases of the experiment: 1) collecting individual preferences and 2) evaluating the recommendations as a group.

*1) **Collecting user preferences**:* A Web System[2] was built for collecting the user preferences. The Web System (see Figure 4) was populated with movie data from the MovieLens dataset [22]. For each movie, we also provided complementary information including *cover*, *imdbLink* and *youtubeID*. The movie covers are listed along with their meta information and the rating option, ranging from 0.5 (dislike) to 5 (like at most). Once a movie is rated, the movie box background becomes green to signalize which movies are evaluated. Each user was asked to rate at least 20 movies so that we could evaluate the approach comprehensively. Worth mentioning that we do not address the cold start problem in this work, but the plans are set for future works.

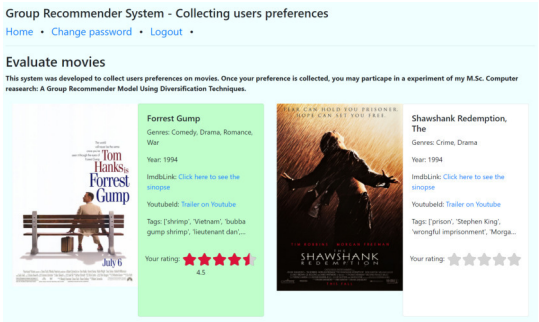

Fig. 4. Collecting users preferences from the Web System.

*2) **Generating recommendations to groups**:* Before generating the recommendations for the groups, we formed the groups. In this work, we did not experiment with any formation heuristics. We simply asked the participants to form groups based on their free will or affinity and only limited the group size to *3* or *5* members. Once the groups were formed, we calculated the group profiles and generated the recommendations. The final group setting was: $G31$, $G32$, $G33$, $G51$, $G52$ and $G53$. For the sake of clarity, we adopt the following terminology *GXY*, where G means *group*, X means the *size* of the group, and Y means the *ID* of the group.

[2]Collecting preferences - https://collectprefgrs.herokuapp.com/

B. Baseline experimental setup

This paper compares our model with the baseline [6]. In their article, baseline's author settled their experiment as the following configuration: 1) They used three datasets, MovieLens, TripAdvisor, and Amazon; 2) Their evaluation metrics were the *S-Recall* and the *Normalized utility*; 3) They formed synthetic groups randomly with $size = 10$ to performed recommendations; 4) They ran a user study were they asked real users to assessed which recommendation list is more diverse, theirs our the baseline ones. Also, they asked the actual users to express which list they preferred.

Therefore, for clarity, in this experiment, we exposed the baseline's algorithm to a new environment, where users are real instead of synthetic. They expressed their ratings for the recommended items. Also, the group size is different, and we assessed using other metrics at different levels.

C. Dataset

MovieLens [22], the dataset used in the experiment, contains 100,000 ratings over 9,000 movies evaluated by 600 users. The original dataset was enriched with data from another experiment with 24 new users and 686 new ratings. Moreover, to each movie, we added it a links to its *cover*, *imdbLink* and *youtubeLink*. We were capable of enhancing the dataset using the *imdbId* information, which is provided in a separate file from the movie's characteristics. This improvement covered approximately 93% of the entire dataset. Despite that, we had to reduce the movie titles' noise to improve the similarity calculations. Table I demonstrates the characteristics of the movie in the final dataset.

D. Metrics

The NDCG, AP, ILD and RMSE were used in the experiment:

*1) **Normalized Discounted Cumulative Gain (NDCG)**:* The NDCG is densely used in Information Retrieval field for measuring the quality of ranked items [23], [24]. The NDCG comprises the value of DCG divided by IDCG. Whereas we recommend the top-k items in a rank, the implementation of the Discounted Cumulative Gain (DCG) and the Ideal Discounted Cumulative Gain (IDCG). The list with top-k items is denoted as $R_k = \{r_1, r_2, ..., r_k\}$, the calculated relevance of an item at position i in the list is represented

by rel_i. The perfect ranking scores 1.0 in this metric. The $DCG@k = \sum_{i=1}^{|R_k|} \frac{rel_i}{log_2(i+1)}$. The $IDCG = max(DCG@k)$ is the maximum value of DCG, and $NDCG@k = \frac{DCG@k}{IDCG}$.

*2) **Average Precision (AP)**:* Precision is the percentage of relevant items recommended to the user [25]. Therefore, $Precision@k$ ($P@k$) represents the percentage of relevant items returned for the user at k level. The $P@k = \frac{RIK}{len(k)}$, where k is the size of the rank to evaluate, RIK stands for **R**elevant **I**tems in the list at **K**, and $len(k)$ is the size of the list at k. In order to observe precision related to the groups size, we implemented the Average Precision (AP) as $AP(G,k) = \frac{\sum_{i=1}^{|G|} P@k}{|G|}$, where G represents a list of groups, and for each group, $P@k$ is considered in the mean for a certain k level for each group in the list. The Precision metric compares the output items with a truth list to set relevance. However, we needed to adapt this metric to evaluate the baseline outputs since they provide no ground-truth list. Thus, inspirited by [17], we defined relevant recommendations as those with a score higher than the global mean of assessed movies.

*3) **Intra List Diversity (ILD)**:* Is a measure for comparing how diversified is a list. The result displays the diversity score, which ranges from 0 to 1, with 0 denoting no diversity and 1 denoting a fully diversified list. The ILD is the antithesis of the Intra List Similarity (ILS) metric, which measures similarity rather than distance between items [15]. We implemented the ILD score in a list of items R as $ILD(R) = \frac{\sum_{i \in R} \sum_{j \in R/\{i\}} Distance(i,j)}{2}$ where $Distance(i,j) = 1 - Similarity(i,j)$. This way, we calculate the mean of all distances from pairs *(i,j)* in list R.

*4) **Root Mean Squared Error (RMSE)**:* The RMSE is a metric used to model error and to help providing a picture of the error distribution [26], [27]. The error considered in this experiment is expressed as $error(r) = (r' - r'')^2$, where for each recommended item r, the truth value r' is subtracted from the predicted value r''. This error function in the RMSE is $RMSE(R) = \sqrt{\frac{1}{len(R)} \cdot \sum_{r \in R} error(r)}$.

E. Results

*1) **Density graph**:* Figure 5 illustrates the rating distribution from user participation. The most common rate is 5.0, with 233 ratings, representing 33.96% of the total assessment. The second most common rate is 4.0, with 134 ratings (19.53%). Worth mentioning that 14.13% of all ratings are under the rate of 3.0. Therefore, the density graph indicates that the group members were pleased with recommendations aligned with the group preference.

*2) **NDCG**:* Figure 6 shows the NDCG results achieved by our approach at positions 3 (nDCG@3), 5 (nDCG@5) and 10 (nDCG@10). On the one hand, the best NDCG result is observed in the group $G33$ with 0.86 at positions 3 and 5. On the other hand, the worst performance is witnessed in the group $G31$, which scored 0.57 at position 5. The other groups varied the NDCG results from 0.72 to 0.80,

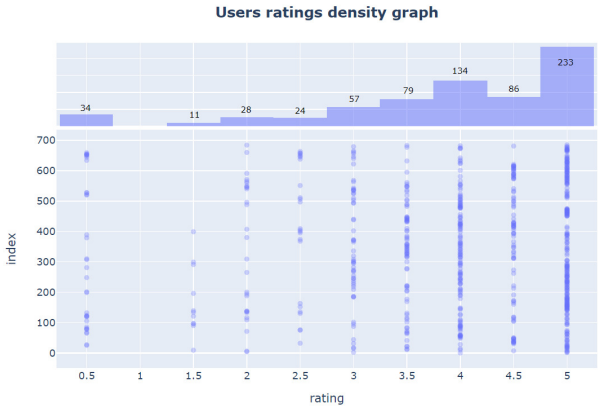

Fig. 5. Density graph of participants' ratings.

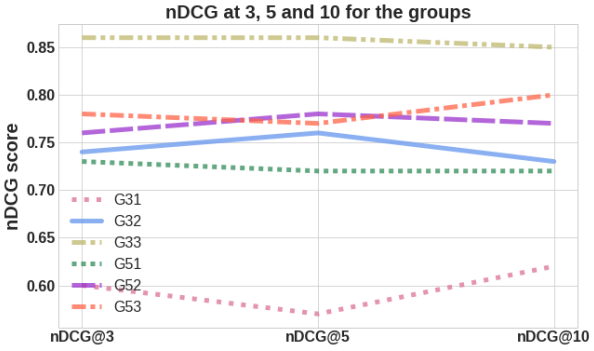

Fig. 6. NDCG for the groups at different levels.

indicating an up-and-coming recommendation list. Looking at the group size, we can observe that the groups with 5 members performed more consistently than the groups with 3 members, especially when we focus on the groups $G31$ (disliked most of the recommendations) and $G33$ (enjoyed most of the recommendations).

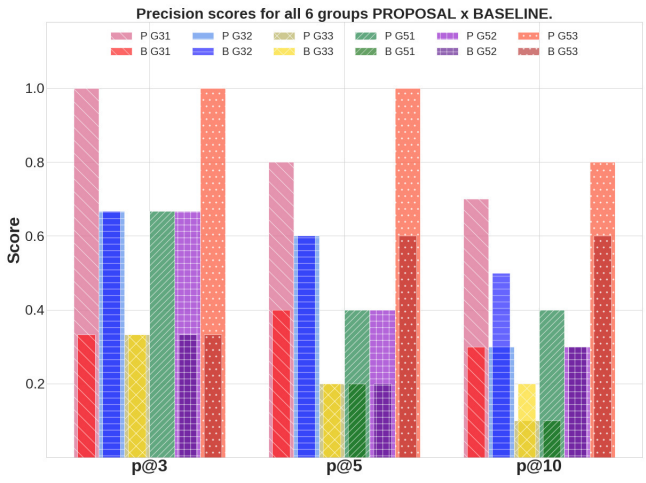

Fig. 7. Precision rates for the groups at different levels. Our proposed model is represented by P and the baseline method is represented by B.

TABLE II
METRICS MATRIX

		nDCG						Average Precision					RMSE
		@3	@5	@10	mean	median	std	@3	@5	@10	median	std	
n=3	G31	0.6	0.57	0.62	0.73	0.74	0.11	0.67	0.53	0.37	0.6	0.20	0.22
	G32	0.74	0.76	0.73									
	G33	0.86	0.86	0.85									
n=5	G51	0.73	0.72	0.72	0.76	0.77	0.03	0.78	0.6	0.5	0.4	0.15	0.20
	G52	0.76	0.78	0.77									
	G53	0.78	0.77	0.8									

3) ***Precision****:* Figure 7 shows the precision results achieved by our approach (P) versus the baseline (B) at positions 3 (p@3), 5 (p@5) and 10 (p@10), for all groups compared. When our approach generates the recommendations, the group $G53$ achieves the highest scores with a surprising precision average of 0.93. The group $G32$ achieves the highest scores with a precision average of 0.58, when the baseline generates the recommendations. As it shows, our approach overcomes the baseline in all compared groups, except for the group $G33$, with the lowest overall result, with an average score of 0.21.

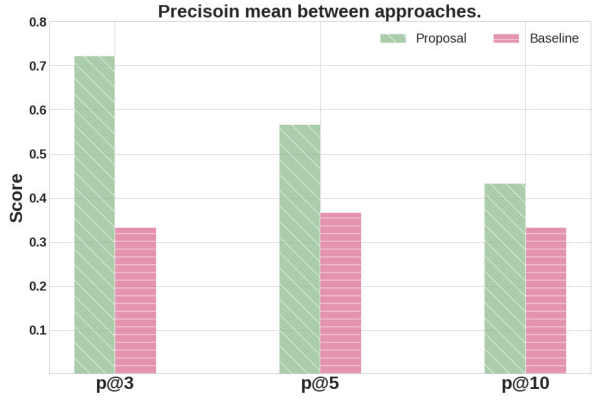

Fig. 8. Precision average between approaches.

4) ***Average Precision****:* Figure 8 shows the average precisions achieved by our approach versus the baseline at positions 3 (p@3), 5 (p@5) and 10 (p@10). The proposed approach overcomes the baseline in all levels, with an advantage of 0.4 points at $p@3$, 0.2 points at $p@5$, and a slight advantage of 0.1 point at $p@10$. In addition, we can observe that as the position increases, the precision means of our approach tend to fall, whereas the baseline seems steady. It is important to point out that the lowest average precision achieved by our approach is 0.44 at position 10, still higher than the highest average precision achieved by the baseline approach, 0.35 at position 5.

5) ***Diversification****:* Figure 9 express the diversity score obtained by the proposal and the baseline. The proposal scores 0.94 with std equals to 0.02, however, the baseline performs better, scoring 0.96 with std equals to 0.01. The baseline method exceeds the proposed one with advantage around 2%.

6) ***Group Size x Metrics****:* Table II summarizes the NDCG, AP, and RMSE results achieved by our approach; however, they are separated by the group size (3 and 5). We can observe

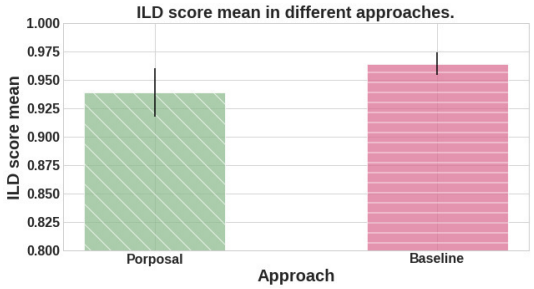

Fig. 9. ILD average score between approaches.

that the groups with 5 members perform better than those with 3 groups for all compared metrics. It is essential to highlight that in the NDCG metric, the std result for the groups with 5 members achieves lower variance than those with 3 members (0.03 against 0.11). In contrast, the opposite occurs for the Average Precision median (0.6 against 0.4). Finally, we evaluated the error rate (RMSE) when filling the sparse matrices. In addition, The RMSE for groups with 3 and 5 members are nearly the same (0.22 against 0.20), achieving a meager error rate.

VI. DISCUSSION AND LIMITATIONS

Despite the promissory results observed, some limitations must be discussed. As to the first research question posed in Section V *i) Are the group recommendations still relevant even after diversification?* The answer is positive, as our approach overcomes the baseline regarding precision. In particular, our approach's lowest Average Precision result overcomes the best Average Precision outcome of the baseline approach. As to the group size, all groups appear to have enjoyed the recommendations, even though the higher variance is observed in groups with 3 members. As to the second research question, *ii) How accurate are the predictions considering the error?* The answer is also positive as the RMSE error rate is approximately 20%, thus leading to relevant group recommendations despite the diversification and group size.

As a limitation, the cold start problem is not treated when no evaluation is observed. A solution can be a hybrid content-based filtering RS. Hence, other metadata such as the actors and director can be compared so that preferences are predicted, and the user matrix is filled out.

Contextual information is another critical aspect that must be carefully incorporated into the proposed model. The motivation that drives a group of people to watch a movie may vary

depending on the occasion and degree of intimacy with the group members, among other aspects. Such an improvement will require designing a more elaborated user and group model that consider such a piece of information.

The diversification addressed in this paper is based on the dissimilarity of key movie features: title and genres. Nevertheless, several other movie characteristics still can be processed, including cast, synopsis, and direction. All these features can be analyzed those can impact diversity.

VII. Conclusion

This paper proposes a group recommendation model that suggests relevant movies for groups based using diversification. The diversification algorithm re-ranks the recommendation list taking into account the relevance of an item and the dissimilarity among them. For evaluating the proposal, a user trial with 24 participants divided into 6 groups was undertaken to assess the proposal. The results show satisfactory results over a baseline method. The proposed approach overcame the compared baseline in all levels of AP evaluation. Moreover, the results point out that the performance for groups of size 5 has a lower variance in std rather than 3.

As for future work, we look forward to performing a deeper study regarding the impact of group size on the recommendations. Also, another online experiment with more participants will help validate the proposal. Additionally, the similarity function can be improved by adding more features related to movies, like directors, cast, or summary. Heuristics for group formation must also be investigated as they impact the acceptance of recommendations. We also plan to test several variations for the diversification algorithm using related metrics MSE and NRMSE. Last but not least, we plan to explain the recommendations to group members. It is already proved that justification serves as essential means to help users to make better decisions among the suggested items.

References

[1] J. Masthoff, *Group Recommender Systems: Combining Individual Models*. Boston, MA: Springer US, 2011, pp. 677–702.

[2] S. Amer-Yahia, S. B. Roy, A. Chawlat, G. Das, and C. Yu, "Group recommendation: Semantics and efficiency," *Proc. VLDB Endow.*, vol. 2, no. 1, pp. 754–765, Aug. 2009. doi: 10.14778/1687627.1687713

[3] A. Jameson and B. Smyth, *Recommendation to Groups*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 596–627.

[4] K. Bradley and B. Smyth, "Improving recommendation diversity," in *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland*. Citeseer, 2001, pp. 85–94.

[5] M. Kaminskas and D. Bridge, "Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems," *ACM Trans. Interact. Intell. Syst.*, vol. 7, no. 1, pp. 2:1–2:42, Dec. 2016. doi: 10.1145/2926720

[6] N. T. Toan, P. T. Cong, N. T. Tam, N. Q. V. Hung, and B. Stantic, "Diversifying group recommendation," *IEEE Access*, vol. 6, pp. 17 776–17 786, 2018. doi: 10.1109/ACCESS.2018.2815740

[7] A. Oliveira and F. Durao, "A group recommendation model using diversification techniques," in *Proceedings of the 54th Hawaii International Conference on System Sciences*, Hawaii, HI, USA, 2021. doi: 10.24251/HICSS.2021.326 p. 2669.

[8] D. Qin, X. Zhou, L. Chen, G. Huang, and Y. Zhang, "Dynamic connection-based social group recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 3, pp. 453–467, 2020. doi: 10.1109/TKDE.2018.2879658

[9] Y. Zheng, "Educational group recommendations by learning group expectations," in *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*, 2019. doi: 10.1109/TALE48000.2019.9225968 pp. 1–7.

[10] D. Karimpour, M. A. Z. Chahooki, and A. Hashemi, "Grouprec: Group recommendation by numerical characteristics of groups in telegram," in *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*, 2021. doi: 10.1109/ICCKE54056.2021.9721494 pp. 115–120.

[11] S. Raza and C. Ding, "A regularized model to trade-off between accuracy and diversity in a news recommender system," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020. doi: 10.1109/BigData50022.2020.9378340 pp. 551–560.

[12] S. Miyamoto, T. Zamami, and H. Yamana, "Improving recommendation diversity across users by reducing frequently recommended items," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018. doi: 10.1109/BigData.2018.8622314 pp. 5392–5394.

[13] D. Bertram, "Likert scales," *Retrieved November*, vol. 2, p. 2013, 2007.

[14] B. Smyth and P. McClave, "Similarity vs. diversity," in *Case-Based Reasoning Research and Development*, D. W. Aha and I. Watson, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001. doi: 10.1007/3-540-44593-5_25 pp. 347–361.

[15] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *Proceedings of the 14th International Conference on World Wide Web*, ser. WWW '05. New York, NY, USA: ACM, 2005. doi: 10.1145/1060745.1060754 p. 22–32.

[16] A. G. Galliano, Ed., *Introdução à sociologia*. Harper e Row do Brasil, 1981.

[17] O. Kaššák, M. Kompan, and M. Bieliková, "Personalized hybrid recommendation for group of users: Top-n multimedia recommender," *Information Processing and Management*, vol. 52, no. 3, pp. 459 – 477, 2016. doi: 10.1016/j.ipm.2015.10.001

[18] R. Ahuja, A. Solanki, and A. Nayyar, "Movie recommender system using k-means clustering and k-nearest neighbor," in *2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, 2019. doi: 10.1109/CONFLUENCE.2019.8776969 pp. 263–268.

[19] S. Girase, D. Mukhopadhyay *et al.*, "Role of matrix factorization model in collaborative filtering algorithm: A survey," *arXiv preprint arXiv:1503.07475*, 2015. doi: 10.48550/arXiv.1503.07475

[20] P. Lops, M. de Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Springer US, 2011, pp. 73–105.

[21] G. Adomavicius, Tuzhilin, and Alexander, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, jun 2005. doi: 10.1109/TKDE.2005.99

[22] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, Dec. 2015. doi: 10.1145/2827872

[23] L. Baltrunas, T. Makcinskas, and F. Ricci, "Group recommendations with rank aggregation and collaborative filtering," in *Proceedings of the fourth ACM conference on Recommender systems*. Barcelona, Spain: ACM, 2010. doi: 10.1145/1864708.1864733 pp. 119–126.

[24] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002. doi: 10.1145/582415.582418

[25] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719, 9780521865715

[26] T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014. doi: 10.5194/gmd-7-1247-2014

[27] E. J. Gilroy, R. M. Hirsch, and T. A. Cohn, "Mean square error of regression-based constituent transport estimates," *Water Resources Research*, vol. 26, no. 9, pp. 2069–2077, 1990. doi: 10.1029/WR026i009p02069