You Do Not Decide for Me!
Evaluating Explainable Group Aggregation Strategies for Tourism

Najafian, Shabnam; Herzog, Daniel; Qui, Sihang; Inel, Oana; Tintarev, Nava

DOI
10.1145/3372923.3404800

Publication date
2020

Document Version
Accepted author manuscript

Published in
Proceedings of the 31st ACM Conference on Hypertext and Social Media (HT '20),

You Do Not Decide for Me!

Evaluating Explainable Group Aggregation Strategies for Tourism

Shabnam Najafian
Delft University of Technology
Delft, the Netherlands
s.najafian@tudelft.nl

Daniel Herzog
Technical University of Munich
Garching bei München, Germany
herzogd@in.tum.de

Sihang Qiu
Delft University of Technology
Delft, the Netherlands
s.qiu-1@tudelft.nl

Oana Inel
Delft University of Technology
Delft, the Netherlands
o.inel@tudelft.nl

Nava Tintarev
Delft University of Technology
Delft, the Netherlands
n.tintarev@tudelft.nl

ABSTRACT

Most recommender systems propose items to individual users. However, in domains such as tourism, people often consume items in *groups* rather than individually. Different individual preferences in such a group can be difficult to resolve, and often compromises need to be made. Social choice strategies can be used to aggregate the preferences of individuals. We evaluated two explainable modified preference aggregation strategies in a between-subject study (n=200), and compared them with two baseline strategies for groups that are also explainable, in two scenarios: high divergence (group members with different travel preferences) and low divergence (group members with similar travel preferences). Generally, all investigated aggregation strategies performed well in terms of perceived individual and group satisfaction and perceived fairness. The results also indicate that participants were sensitive to a dictator-based strategy, which affected both their individual and group satisfaction negatively (compared to the other strategies).

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Human-centered computing** → **User studies**; **Empirical studies in HCI**.

KEYWORDS

Group recommendation; Explainable aggregation strategies; Human-centered computing user studies

ACM Reference Format:
Shabnam Najafian, Daniel Herzog, Sihang Qiu, Oana Inel, and Nava Tintarev. 2020. You Do Not Decide for Me!: Evaluating Explainable Group Aggregation Strategies for Tourism. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media (HT '20), July 13–15, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3372923.3404800

HT '20, July 13–15, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7098-1/20/07…$15.00
https://doi.org/10.1145/3372923.3404800

1 INTRODUCTION

Recommender systems can help users to cope with an abundance of items to try or buy by offering those items the user is likely to find interesting [4]. However, in domains such as tourism people often consume items in groups rather than individually, creating a need for strategies for Group recommender systems.

Previous work suggests strategies for combining users' individual preference models into an aggregated group prediction [19]. However, there is no optimal way to do this; every feasible aggregation method has some disadvantages [2]. Besides, different aggregation strategies may be effective for different situations and different kinds of groups. For example, items with higher average ratings are not good recommendations when the people in the group have very different preferences.

To mitigate the disadvantages of aggregation strategies and avail their advantages, two new explainable aggregation strategies have been proposed by combining existing aggregation strategies [22]. However, to the best of our knowledge, these two aggregation strategies have not yet been evaluated in real group tourism recommendation scenarios. Therefore, in this work, we evaluate these modified aggregation strategies in terms of their ability to recommend sets of points of interest (POIs) for groups of tourists. We focus specifically on situations where group members have different travel preferences (*high divergence*) versus similar travel preferences (*low divergence*). We assess the impact of the modified aggregation strategies proposed in [22] by comparing them to the `Average` and `Dictatorship` strategies as our baseline strategies in an online study. Specifically, we investigate the following research question:

Which strategy performs the best in which scenario (level of divergence) in terms of user perceived individual and group satisfaction, perceived fairness, and user acceptance?

In summary, in this paper we make the following contributions:

- We investigate which of four explainable aggregation strategies help increase user-perceived satisfaction, fairness, and acceptance.
- We investigate whether the level of divergence in group members' travel preferences influences user-perceived satisfaction, fairness, and acceptance.

The next section outlines related work in preference aggregation strategies for recommending items to groups and different

approaches to calculate the level of preference divergence in a group. In Section 3, we elaborate on the four aggregation strategies that will be further evaluated in a user study. In Section 4, we describe the user study performed to evaluate the proposed aggregation strategies in terms of individual and group satisfaction, fairness, and acceptance. Then, we describe the results and discuss the limitations of our study in Section 5. We conclude with plans for future work in Section 6.

2 RELATED WORK

In this section we introduce related work on so-called aggregation strategies which allow us to combine individual preferences for multiple items and users. Since group members can have different preferences, the resulting recommendations may result in compromises. Therefore, we have chosen to focus on aggregation strategies that can be explainable and intelligible to individual group members.

In this section we also describe ways of defining differences in user preferences for group recommendation.

2.1 Aggregation strategies

There are a number of alternative approaches to aggregating preferences in group recommender systems, for instance [13, 14]. Group recommendations typically can be generated either by aggregating predictions or aggregating models [10]. In this paper we focus on the former, which determines items/ratings for individual group members, and thereafter aggregates these items/ratings to a group recommendation.

Table 1 provides an overview of the different strategies considered in this paper and inspired by Social Choice Theory (deciding what is best for a group given the opinions of individuals). Further strategies can be found in Chevaleyre et al. [6], Masthoff [18, 19], Senot et al. [25]. As stated by Arrow's theorem [2], there is no aggregation strategy that outperforms all the other strategies in all situations, suggesting the need to evaluate when different strategies are most useful.

The most frequently used aggregation strategy in the literature to provide recommended items for group members is the `Average` strategy ([1], [3], [12], [21], and [24]). Although this appears reasonable, this method does not always guarantee high quality recommendations for groups because it is unable to reflect the propensities of all users in a group [26]. For example, if a group has highly divergent preferences, then using the average increases the number of recommended items that members do not prefer (as long as enough other members do).

Previous work evaluating aggregation strategies with users has inquired which aggregation strategies were employed by real people [17], and found that people use strategies such as the *Average Strategy*. The subjects were presented with an example of ratings by 3 synthetic users for 10 specific items (rated by all users). From these ratings they were asked to generate a sequence of items as a group.

Herzog and Wörndl [15] conducted an online study where all group members shared a public display. The group had to agree verbally on group preferences and enter them into the system to receive a recommendation. Their results show that the preferences entered on the public display resembled mostly to the `Average` strategy (52.5% of groups). The second most applied strategy was the `Dictatorship` strategy (39.4% of groups). Therefore, we picked these two strategies as our baseline strategies.

Previous work has also suggested new aggregation strategies [22] that aim to improve upon existing ones. However, to the best of our knowledge, these aggregation strategies have not yet been evaluated. Therefore, in this study we evaluate two modified strategies, namely `Least+` and `Fair+`, as proposed in [22] and compare them to our two baseline strategies. In addition, we study the influence of group members' travel preferences divergence on the performance of each strategy.

2.2 Level of preference divergence in a group

Different approaches to calculate the similarity or distance between group members' preferences exist. We refer to the similarity distance in this paper as the level of divergence between group members' preferences.

For example, Herzog and Wörndl [15] applied the Pearson correlation coefficient (PCC), which referred to as Pearson's r, to determine the similarity of two group members' travel preferences. Delic et al. [8] used the Full Choice-set Distance measure (FullDist). It considers members' preferences for the full set of options (ChoiceSet), to compute the distance between two group members u and v. It gives an undirected preferences relationship between pairs of group members:

$$FullDist(u, v) = \sum_{i \in ChoiceSet} |score_u(i) - score_v(i)| \qquad (1)$$

Seo et al. [26] measured the mean square deviation (MSD) to calculate the deviation of preference ratings for items in groups. The value of MSD, shows the distribution of ratings for each item.

In this paper the level of divergence refers to how much people in the group have different or similar travel preferences. For this purpose we use Pearson's r which measures the linear correlation between two variables and is often used in RSs to identify similar users [16].

3 AGGREGATION STRATEGIES

In this section we describe the four aggregation strategies that we compare and evaluate in our user study: `Least+` and `Fair+`, proposed in [22], `Average` and `Dictatorship`, proposed in [17].

We describe each of these aggregation strategies with examples (from [17]) with individual ratings for 10 items (A to J) for a group of three (John, Adam, and Mary). The highest possible rating is 10. The Sum row calculates the final scores for each item. Group List represents the sorted final recommended list.

`Least+` *(Least Misery + Most Pleasure + Without Misery) [22].* The `Least+` strategy prioritizes, and presents first, items that maximize the rating of the happiest person and at the same time minimize the unhappiness of the saddest person within the group [22]. The *Most Pleasure* strategy considers the highest rating in the group as a group preference rating for the item. The *Least Misery* strategy means that the preferences of a group to items are decided by the lowest rating in the group (the least happy member). The *Least Misery* strategy is one of the prevalent ones and it has been widely

Table 1: Applied aggregation strategies in this paper. Fair+ and Least+ have not previously been evaluated in user studies.

Aggregation strategy	Description	Disadvantage
Average	Average of individual ratings	It does not consider extreme cases, and it is not optimal method when the individual preferences highly diverge because e.g. extreme low ratings can be balanced out by extreme high ratings.
Fairness	Item ranking as if individuals (u ∈ G) choose them one after the other	It does not consider low ratings if it is top item of an individual.
Least Misery	Minimum of individual ratings	A minority opinion can dictate the group, it only considers minimum ratings in group.
Most Pleasure	Maximum of individual ratings	It does not consider low ratings, also so many ties could occur when the scope of rating is fixed, so many items could have the same score.
Dictatorship	Rating of most respected individual	A minority opinion can dictate the group, it only considers the opinion of the most respected person in the group.
Without misery	Avoiding low ratings	A minority opinion can dictate the group, applies a veto to an item rated below a threshold by any user.
Fair+	*Fairness* → *Average* (Combination of strategies)	May include an average rated item if it is top item of one individual.
Least+	Least Misery + Most Pleasure + Without Misery (Combination of strategies)	A minority (negative) opinion can dictate the group.

applied in traditional group recommender approaches [12]. The *Without Misery* strategy excludes items that anyone in the group rated below a certain threshold. When using the original *Least Misery* and *Without Misery* strategies on their own, items may be selected such that nobody dislikes, but also, nobody really likes. An example of the `Least+` strategy can be seen in Table 2. The LM row shows the items' scores after applying the *Least Misery* strategy. This strategy makes a new list of ratings with the minimum of the individual ratings for each item. The next row, MP, shows the items' scores after applying the *Most Pleasure* strategy. This strategy makes a new list of ratings with the maximum of the individual ratings for each item. Finally, the Sum row shows the sum of LM and MP rows. The dashes in each row indicate that the item will not be considered for recommendations.

`Fair+` *(Fairness -> Average) [22].* The `Fair+` strategy takes turns between users to select their most preferred item, which corresponds to the item with the highest ranking in the rated items list for users. This strategy considers the satisfaction of all the users but could include the most hated item if it is a top item of one member. This strategy in group settings can be characterized as a strategy without favoritism or discrimination towards specific group members [11], compared to `Least+`, where one member could dictate her preferences. In the `Fair+` strategy, one person chooses first, then another, until everyone has made one choice. The next rounds begin with the one who had to choose last in the previous round. When the rating is the same for multiple items, the item with the higher average rating will be selected. An example can be seen in Table 3. In our example, if we start with John first, his favorite items are A, E, or I. We recommend E because it has the highest average. Next, it is Adam's turn. Adam would like B, D, F, or H. We recommend F because it has the highest average. Mary would choose A (her highest rating). Next, we start with Mary, she would like E, which has already been recommended, and then F, which also has already been recommended. Following the Masthoff [17] approach, we then skip Mary's preferences in this round and recommend based on Adam's highest rating. He likes B, D, or H. We recommend H, as that has the highest average. Following this strategy, we could end up with a group list like: E, F, A, H, I, D, B, J, C, G.

Table 2: Applying the `Least+` (*Least Misery (LM) + Most Pleasure (MP) + Without Misery (WM)*) strategy on an example from [20]. LM and MP rows show the items' scores after applying the LM and the MP strategies respectively. The dashes in each row shows that item will not be considered for recommendations because of the WM strategy.

	A	B	C	D	E	F	G	H	I	J
John	10	4	3	6	10	9	6	8	10	8
Adam	1	9	8	9	7	9	6	9	3	8
Mary	10	5	2	7	9	8	5	6	7	6
LM	-	4	-	6	7	8	5	6	-	6
MP	-	9	-	9	10	9	6	9	-	8
Sum	-	13	-	15	17	17	11	15	-	14

Group List: (E, F), (H, D), J, B, G (threshold 3 out of 10)

`Average` *[17].* This strategy averages individual ratings and selects items with high average ratings. It does not consider extreme cases, and it is not an optimal method when the individual preferences highly diverge because, for example, extreme low ratings can be balanced out by extreme high ratings. An example can be seen in Table 4.

Table 3: Applying the `Fair+` *(Fairness -> Average)* strategy on an example from [20]. For the sake of readability the sum is not divided by number of group members.

	A	B	C	D	E	F	G	H	I	J
John	10	4	3	6	10	9	6	8	10	8
Adam	1	9	8	9	7	9	6	9	3	8
Mary	10	5	2	7	9	8	5	6	7	6
Sum	21	18	13	22	26	26	17	23	20	22

Group List: E, F, A, H, I, D, B, J, C, G

Table 4: Applying the *Average* strategy on an example from [20]. For the sake of readability the sum is not divide by the number of group members.

	A	B	C	D	E	F	G	H	I	J
John	10	4	3	6	10	9	6	8	10	8
Adam	1	9	8	9	7	9	6	9	3	8
Mary	10	5	2	7	9	8	5	6	7	6
Sum	21	18	13	22	26	26	17	23	20	22

Group List: (E, F), H, (J, D), A, I, B, G, C

`Dictatorship` *[17].* In the Dictatorship strategy (also called 'Most Respected Person strategy'), only the ratings of one member in the group will be considered for generating the recommendations to the group. In this strategy, the group may be dominated by one person. For example, if you respect highly a person in the group, like your boss, you may all follow his/her taste. An example can be seen in Table 5. In our case we always selected one of the other group members' preferences rather than the active user.

Table 5: Applying the *Dictatorship (Most Respected Person)* strategy on an example from [20]. In this example, the ratings of Adam are considered as a dictator.

	A	B	C	D	E	F	G	H	I	J
John	10	4	3	6	10	9	6	8	10	8
Adam	1	9	8	9	7	9	6	9	3	8
Mary	10	5	2	7	9	8	5	6	7	6

Group List: (B, D, F, H), (C, J), E, G, I, A

Summary of strategies and their trade-offs. The `Average` and the `Dictatorship` strategies serve as baselines as they have been the most applied strategies by groups (Herzog and Wörndl [15]). Besides, we will evaluate `Least+` and `Fair+` strategies. We believe that it is interesting to compare these two strategies (`Least+` and `Fair+`) as they have complementary strengths and weaknesses. In one (`Least+`), having high average satisfaction by excluding the least preferred item(s) of one or more people. In the other (`Fair+`), having a fair system that might recommend to you your most hated item if it is a top item of another group member (as long as you get to visit the places you really love as well).

4 USER STUDY

We wanted to understand which of the previously introduced strategies performs better in terms of perceived satisfaction, fairness, and acceptance in high divergence scenarios and low divergence scenarios. For this purpose, we recruited crowd-workers from Amazon Mechanical Turk (MTurk)[1] to conduct a user study.

4.1 Preliminaries

4.1.1 Data Set. Our first task was to compose a set of recommendations for the user study. We use preferences for different categories from a previous travel-related user study [15]. In that study, every user individually rated 42 categories (e.g., *Art Museum* and *French Restaurant*), on a scale from 0 (not interested in this category) to 5 (strongly interested in this category). There were 40 groups with 3 members registered for the study. The groups were real, i.e. participants applied as groups and were not randomly assigned. The participants were asked to imagine the scenario "single-day trip in Munich".

4.1.2 Selecting items to rate. To obtain the crowd workers' preferences we wanted to provide them with an initial 42 POIs to rate. We retrieved the most popular POI (in terms of like count) for each selected category (for all 42 categories from the data set) from the social location service Foursquare[2] as a representative of that category. By using a real data set we increased the likelihood of a realistic rating distribution.

4.1.3 Group composition. In our experiment, we want to force high divergence and low divergence in the group. To do this, based on crowd workers' ratings for the 42 initial POIs, we form a group for the crowd worker by picking two synthetic group members from the real tourist data set that we described in the Section 4.1.1. For half of the crowd workers, we select two users with the **highest similarity** compared to the crowd worker and for the other half two users with the **lowest similarity** (dissimilar). We did not consider how similar or dissimilar the other two synthetic group members are to each other, as we were not interested in the average group divergence, but we were only interested in the level of divergence towards the real user. A user's travel preferences are represented by a vector of length 42. We used the Pearson's r to determine the similarity/dissimilarity of two user's travel preferences (the potential options are discussed in Section 2.2).

4.2 Independent variables

We manipulate the following (independent) variables in this study:

Aggregation Strategies. `Least+` and `Fair+` as modified strategies as well as `Average` and `Dictatorship` as baseline strategies.

Levels of Divergence. In this study we consider two levels of divergence: `high divergence` and `low divergence`. We believe it is more important to study `high divergence` cases because it is more challenging to satisfy all group members when they have different travel preferences. For the sake of comparability we also applied strategies on the groups we predicted to have a low divergence in their preferences and have more similar travel

[1] https://www.mturk.com, retrieved November 2019.
[2] https://developer.foursquare.com/, retrieved April 2019

Figure 1: Screenshots of the task. First, participants see the left page which includes: (A) input fields to ask participant to enter the names of the imagined group members, (B) the 42 initial POIs to obtain participant's preferences in the scenario "Single-day Trip in Munich". Then, participants see the right page which includes: (C) recommended POIs generated by one of the four strategies, (D) the description of the scenario and the explanations of how the strategy works and ratings of the recommended POIs given by the participant and other two group members, and (E) questions for evaluating the recommended POIs.

preferences. As explained in Section 4.1.3, we calculated Pearson's r between group members within the group. The range of values for Pearson's r is between -1.0 to 1.0, where -1.0 indicates the strongest negative correlation of travel preferences of two users (contrary preferences) and 1.0 indicates the strongest positive correlation of travel preferences of two users (similar preferences). We consider values between [-1, 0) as `high divergence` and values between (0, 1] as `low divergence`.

4.3 Dependent variables

Each recommended POIs list was evaluated according to four dimensions: perceived individual and group satisfaction, perceived fairness and user acceptance. For this purpose, each participant received the following questions on a 7 point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree):

Perceived individual satisfaction: *"I'm satisfied with the recommended places to visit."*

Perceived group satisfaction: *"I believe the recommended places to visit are good recommendations for this group."*

Perceived fairness: *"I believe the recommended places to visit represent every group members' interests in the group."*

We also asked the users to give us the number of places they would like to visit from the recommended list, *i.e.*, the user acceptance. There is a total number of 6 POIs that we recommend.

User acceptance: *"How many places from the recommended places would you visit?"*

Finally, a free-text comment as the last question was provided for participants to motivate their answers.

4.4 Procedure

We designed an online between-subjects experiment in which participants were randomly assigned into a 4 strategies (`Least+`, `Fair+`, `Average`, and `Dictatorship`) × 2 levels of divergence (low vs high) design. We created 8 = 4 × 2 different variations, manipulating strategies and levels of divergence, and each participant only sees one variation. This allows us to evaluate the aggregation strategies in terms of perceived individual and group satisfaction, fairness, and user acceptance in two different scenarios (low divergence vs high divergence). The measurement is the same for all 8 variations as defined in Section 4.3.

The experiment consisted of three steps (see Figure 1):

Step 1: Some of the participants' individual details were collected, such as demographics (age, gender), education level, and frequency of using apps for recommending touristic places.

Step 2: Next, they were asked to imagine the scenario "single-day trip in Munich" and rate the 42 predefined POIs (see Section 4.1.2). The POIs were augmented by Google Street View for a more accurate preference rating (see Figure 1 (B)).

Step 3: According to a participant's initial ratings, for half of the crowd workers a high divergence group and for the other half a low divergence group is created (see Section 4.1.3). Next, a list of POIs is generated, based on all group members' preferences by applying randomly one of the four aggregation strategies, namely: `Least+` (we apply threshold 2 out of 6 in our experiment), `Fair+`, `Average`, and `Dictatorship` (we recall here that for the `Dictatorship` strategy, we always selected one of the other group members' preferences rather than the active user). We presented the top 6 POIs from the generated recommendations since it was a more realistic

length for a one-day touristic visit. The recommended POIs were presented as a set, and participants were told that, "this set does not contain any order and can be consumed in any order".

They were presented with explanations of how the strategy works and what the other two group members' ratings are. Figure 1 (C, D), illustrates the recommended POIs and its corresponding explanations. Participants can additionally explore each POI using Google Street View.

We asked participants to answer a set of survey questions related to evaluating the recommended POIs in terms of perceived individual and group satisfaction, perceived fairness, and user acceptance (see Section 4.3). We also included the following attention check question: "This is an attention check. Choose Disagree" to exclude malicious participants. At the end of the survey, participants were asked to express their opinions regarding the recommended POIs and recommendation algorithm in an open-ended question (see Figure 1 (E)).

4.5 Hypotheses

In the following, we refer to perceived fairness, acceptance, individual satisfaction, and group satisfaction as F-A-IS-GS to avoid making hypotheses long and hard to differentiate.

Given the trade-off between the strategies and the level of divergence, we hypothesize that:

- **H1)** F-A-IS-GS vary across different strategies.
- **H2)** F-A-IS-GS differ in levels of group preference divergence.
- **H3)** F-A-IS-GS vary across different strategies and levels of group preference divergence.

4.6 Statistical analyses

We wanted to determine if there are non-random associations (with regard to the dependent variables) between eight categorical variables (in our case four strategies and two levels of divergence).

To test our hypotheses, we applied the Two-way MANOVA test for between-subjects. Bonferroni correction was applied when multiple tests were conducted. The required sample size was estimated to be 200 participants, this was based on the G*Power analysis for the Two-way MANOVA for between-subjects user studies [9].

5 RESULTS

In this section we describe the results of evaluating four different preference aggregation strategies to recommend POIs in the context of groups, in two scenarios, namely high divergence scenarios and low divergence scenarios.

Participants. We recruited 226 participants from MTurk in December 2019. All participants are based in the United States and have overall HIT approval rates of at least 95%. Knowing Munich was not a pre-requisite for the study. Each participant received $2 as compensation for their time (on average it took 21 minutes of their time to complete the entire task). We excluded 26 participants who failed the attention check question from our data analysis. This resulted in 200 participants (50 per strategy): 38.5% female and 61.5% male. The highest level of education that they held was 29% a high school diploma or equivalent degree, 57% a bachelor's, and 29% a master's degree or higher. Among those, 32.5% use tourism applications (such as Yelp, or Foursquare) less than once a month, 29% at least every month, 22.5% every week and only 8% never used one.

Table 6: MANOVA: Wilks Test – it tests the main effect between the strategies, between the levels of divergence and the interaction between the strategies and the levels of divergence on the combined dependent variables

Cases	df	Approx. F	Wilks' Λ	Num df	Den df	p
(Intercept)	1	2255.045	0.022	4	203.000	< .001
strategy	3	2.388	0.872	12	537.379	**0.005**
divergence	1	2.097	0.960	4	203.000	0.083
strategy * divergence	3	0.558	0.968	12	537.379	0.876
Residuals	200					

5.1 H1: F-A-IS-GS vary across different strategies

There was a statistically significant main effect between strategies on the combined dependent variables (F = 2.390, p = .005; Wilks' Λ = .871) (see Table 6). We did between-subjects effects test (ANOVA) to investigate further the effect on each dependent variable. Tests of the four hypotheses (four variables) were conducted using Bonferroni adjusted alpha levels of .0125 per test (.05/4). Following we discuss its results.

H1.1: in terms of perceived individual satisfaction (IS). The results showed there is a significant difference in perception of individual satisfaction between strategies (p < .0125). The post hoc test showed the three strategies (`Least+`, `Fair+`, and `Average`) have significantly higher perceived individual satisfaction compared to the `Dictatorship` strategy (p < .0125) (see Table 7 for the average and standard deviation values).

H1.2: in terms of perceived group satisfaction (GS). The results showed there is a significant difference in perception of group satisfaction between strategies (p < .0125). We did post hoc test to see which strategy led to higher group satisfaction. The results showed that the `Fair+` strategy has significantly higher perceived group satisfaction compared to the `Dictatorship` strategy (p < .0125) (see Table 7 for the average and standard deviation values).

H1.3: in terms of perceived fairness (F). There was no significant difference between the strategies in terms of perceived fairness.

H1.4: in terms of user acceptance (A). There was no significant difference between the strategies in terms of user acceptance.

5.2 H2: F-A-IS-GS differ in levels of group preference divergence

There was no statistically significant main effect between different levels of divergence on the combined dependent variables (F = 2.097, p = .083; Wilks' Λ = .960). Therefore, all the sub hypotheses for each individual variable are rejected accordingly. This result will be discussed further in Section 5.5.

Table 7: The table shows the average and deviations of the study results for the user perceived individual satisfaction, group satisfaction, fairness, and acceptance. The three first variables are evaluated by 7 point Likert scale, and acceptance shows how many POIs among 6 recommended POIs user would accept. The maximum values for the four measured variables are in bold.

		Individual Satisfaction		Group Satisfaction		Fairness		Acceptance	
Strategy	**Divergence**	**Mean (out of 7)**	**Std**	**Mean (out of 7)**	**Std**	**Mean (out of 7)**	**Std**	**Mean (out of 6)**	**Std**
Least+	Low	5.73	1.07	5.69	0.85	5.34	1.01	2.85	0.46
	High	5.66	1.23	5.46	1.39	5.15	1.61	2.62	0.67
	Total	**5.69**	1.15	5.58	1.16	5.25	1.32	2.73	0.59
Fair+	Low	5.62	1.23	**5.76**	1.21	5.45	1.27	2.92	0.39
	High	5.48	1.32	5.73	1.00	5.27	1.21	2.76	0.63
	Total	5.55	1.27	5.75	1.10	5.36	1.23	2.84	0.53
Ave	Low	5.68	0.85	5.78	1.12	5.52	0.79	2.92	0.27
	High	5.61	1.30	5.44	1.00	5.28	0.98	2.91	0.28
	Total	5.65	1.08	5.60	1.06	**5.40**	0.89	**2.92**	0.27
Dict	Low	4.97	1.47	5.10	1.11	5.10	1.32	2.83	0.37
	High	4.62	1.72	4.88	1.53	4.85	1.82	2.69	0.55
	Total	4.80	1.59	5.11	1.33	4.98	1.57	2.77	0.47
Total	Low	5.48	1.22	5.53	1.36	5.29	1.29	2.74	0.57
	High	5.35	1.45	5.48	1.00	5.20	1.29	2.88	0.38
	Total	5.41	1.34	5.50	1.19	5.24	1.29	2.81	0.49

5.3 H3: F-A-IS-GS vary across different strategies and levels of group preference divergence

There was no statistically significant interaction effect between strategies and the level of divergence on the combined dependent variables ($F = 0.558$, $p = .87$; Wilks' $\Lambda = .968$). Therefore, accordingly, all the sub hypotheses for each individual variable are rejected. This result will be discussed further in Section 5.5.

5.4 Post hoc analysis

All aggregation strategies performed well in terms of all defined dependent variables. There is a difference between strategy performance but the difference is smaller than we expected. We investigated which factors contributed to the surprising results. We particularly looked at differentiation between strategies, the number of common items between strategies, and differentiation between levels of divergence.

5.4.1 Lack of differentiation between strategies. We checked whether we captured the weakness of the applied strategies.

The `Least+` *strategy did not exclude the most favorite item from the recommended set.* The main weakness of the `Least+` strategy is excluding the highly rated item from the recommended set if it is below a certain threshold for another group member. The `Least+` strategy excluded the top most favorite item for only 6/50 participants.

The `Fair+` *strategy did not include the least favorite item in the recommended set.* The Fair+ strategy has a weakness that it may include a least rated item of a group member if it is a top item of another group member. We checked how often this happened in this study, and it only occurred for 4/50 participants.

The `Average` *strategy did not have extreme low ratings or extreme high ratings.* The main weaknesses of the `Average` strategy is that it does not consider extreme cases, and it is not an optimal method when the individual preferences highly diverge because, e.g., extremely low ratings can be balanced out by extremely high ratings. As can be seen in Figure 5, our study did not contain very high divergence groups.

The `Dictatorship` *strategy recommended items that represent all group members' preferences and not only the preferences of one member.* The main weakness of the `Dictatorship` strategy is that it only considers and recommends based on one group member's preferences. Moreover, it is not an optimal method when the individual preferences are highly divergent. In our set-up, always a member other than the active user dictates her preferences. In Figure 2 we show the active user's ratings for each POI, recommended by each strategy. The results show that the average ratings of the active user to the recommended POIs were lower for the `Dictatorship` strategy, compared to the average values of the other three strategies. This can also motivate the difference we found between Dictatorship and other three strategies in terms of individual and group satisfaction (both values were lower for the Dictatorship strategy).

5.4.2 Number of common items between strategies was high. To understand the similar results between the strategies, we checked to what extent strategies recommended similar or different POIs. In Figure 3 we see that three-quarters of the total amount of POIs, namely 30 out of 42, are in common among all four strategies (10

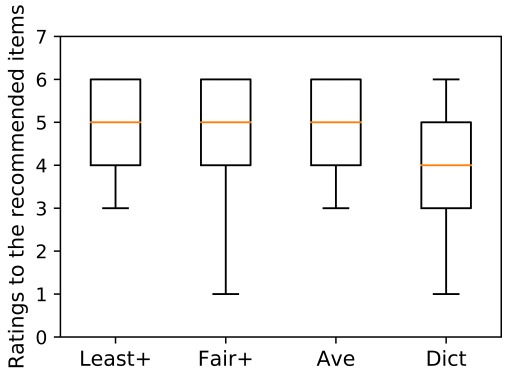

Figure 2: The initial ratings (range [1,6]) of the active user for the recommended POIs by each strategy.

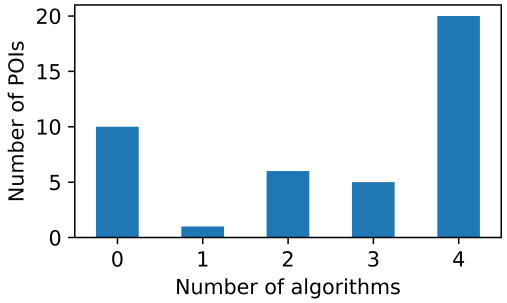

Figure 3: Number of POIs recommended by none, one, two, three or all four strategies.

POIs have never been recommended in any strategy and 20 POIs have been recommended in all 4 strategies). Only 12 POIs have been recommended either by one, two or three of the strategies. Besides, we looked at how different was the behavior of strategies. Figure 4 shows the pairwise comparison of occurrence of each POI between each pair of strategies. It can be seen that the median of differences for all pairs of strategies is below three (which is a small number). It shows that, in general, strategies did not recommend very different POIs.

5.4.3 Limited differentiation between levels of divergence. We did not find a significant effect of scenario (high and low divergence groups) in terms of users' perceived satisfaction, fairness, and acceptance. We investigated whether this was due to a lack of differentiation between these groups.

Figure 5 shows the box plot of Pearson's r values of a user-user pair in a group (two values per group, active user vs acquaintance 1 and active user vs acquaintance 2).

As can be seen in the Figure, there is a differentiation between high and low divergence groups. However, the median correlation for groups with high divergence is not very low (around -0.12), meaning that, overall, we did not have very strong divergence groups.

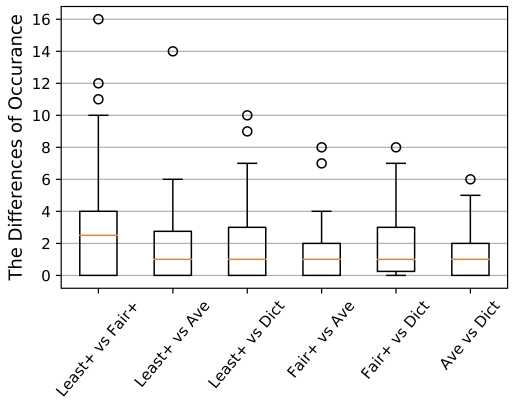

Figure 4: Pairwise comparison of frequency of occurrence of all 42 predefined POIs between strategies: Least+ vs Fair+, Least+ vs Average, Least+ vs Dictatorship, Fair+ vs Average, Fair+ vs Dictatorship, Average vs Dictatorship.

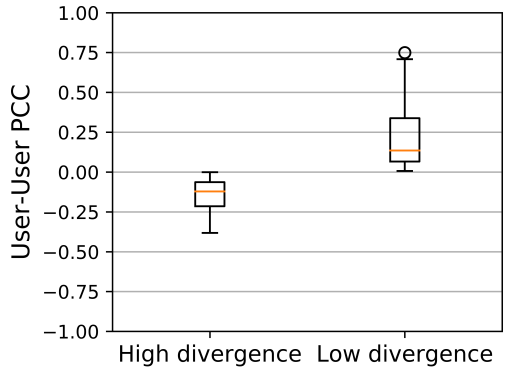

Figure 5: The distribution of similarity of travel preferences between users (active user and two non-active users) in high divergence groups and low divergence groups. The similarity is computed using Pearson's r. In our experiments, the similarity of travel preference for high divergence groups is between [-0.40, 0] and for low divergence groups is between [0, 0.75].

Rather than ensuring divergence on the candidate POIs in the user profiles, we could have enforced diversity in the resulting recommendation list. For example, using a metric such as Intra List Distance (ILD) [27, 28].

However, it can be the case that having a very high divergence is an artificial scenario, especially in the tourism domain where recommended POI are often popular (and liked by many group members). We could have used completely synthetic data where we fixed the divergence levels, but this would decrease ecological validity.

5.5 Discussion

Despite the differences in `Least+`, `Fair+`, and `Average`, both the perceived individual and group satisfaction is comparable and high for all three strategies. However, these three strategies have significantly higher individual and group satisfaction compared to the `Dictatorship` strategy. Some user comments indicated that they are not fully satisfied with the recommended set because it only considered one person's preferences (boss's preferences) in the group.

This can be interpreted as follows: people are more sensitive when the strategy represents the preferences of one member in the group and does not consider other members.

Given the similar results for the strategies and scenarios, we identified in post-hoc analysis that the following factors may have contributed to our results.

Lack of differentiation between strategies. Given that the weaknesses of the different strategies did not happen often, they were all given high ratings.

Number of common items between strategies was high. Given that the strategies often recommended the same POIs, it might be not so important which strategy one applies, although this is likely to depend on rating distributions, domain, among others.

Lack of differentiation between two levels of divergence. While the two levels of divergence were distinct, this difference could have been stronger.

5.5.1 Limitations. The online experiment used real POIs and user data, which allowed us to have a more realistic scenario, especially by using street view in a European city. On the other hand, using previous ratings constrained our ability to emphasize the differences in scenarios or strategies as might be done in a controlled (but possibly implausible) setting.

Our study only measured the evaluation of one member in a group, when the dictator was someone else in the group, and someone they respected.

6 CONCLUSION AND FUTURE WORK

In this section, we highlight the final conclusions and plans for future studies.

6.1 Conclusion

In this paper we presented a user evaluation of four different explainable aggregation strategies, namely `Least+`, `Fair+`, `Average`, and `Dictatorship`, in two scenarios (groups with different preferences versus groups with similar preferences) in the *tourism* domain. We found a significant difference between algorithms in terms of the combined variables (perceived individual and group satisfaction, fairness and acceptance). Further analysis showed a difference between the `Dictatorship` strategy and the other three strategies in terms of both user perceived individual and group satisfaction. User comments suggest that our participants were sensitive to the dictator-based strategy which (comparatively, negatively) affected their satisfaction on their own behalf, as well as on behalf of the group.

We also saw that all strategies performed well in terms of both individual and group satisfaction. This suggests that at least in this setting and domain, they may all be suitable for group recommendations.

6.2 Future directions

This work has also highlighted a number of exciting avenues for future research.

Effects of level of divergence. Surprisingly, we did not find an effect of scenario (whether group members have different versus similar preferences). This may be due to a constraint in terms of the divergence in the data sample used. Further work is required to investigate whether this is inherent to groups who make decisions together, the tourism domain, or this specific dataset.

Effects of aggregation strategies. In addition to the points we mentioned in the paper, there can be other reasons for explaining the individual satisfaction or dissatisfaction regarding the recommended POIs. For the next stage, we plan to take inspiration from the Technology Acceptance Model (TAM) [7] and its extended models, in order to better understand the user responses (e.g., integrating group dynamics such as relationship type within the group, group composition (e.g. minority vs majority, etc.), and more general ones, such as recommendation diversity [23], user's attitudes [23], among others). Our study only measured the evaluation of one member in a group, when the dictator was someone else in the group, and someone they respected. Further work is planned to evaluate the effect of satisfaction of recommendations given by a `Dictatorship` strategy for different members in a group (e.g., based on profile similarity or their relationship).

Generalizability. To better understand the generalizability of our results we will study the effect in real groups, as well as conduct studies in a different domain (music).

Explainable aggregation strategies. The fact that these simple strategies are effective creates a foundation for further research on explainable group recommendations. For example, an explanation for the `Fair+` strategy could be as follows: *"The system detected you might not like the recommended Van Gogh museum but it is the museum that Bob prefers the most. You made your choice in the previous round, now it's Bob's turn to pick! Would you like to reconsider?"*

In addition to the aggregation strategies evaluated in this study, there are other alternative strategies that could be explored. For instance, Carvalho and Macedo [5] introduced game theory into group recommendation and transformed the recommendation problem into finding the Nash equilibrium. There was no empirical evaluation of these methods with people in the groups and no explanation has been designed yet. It needs more exploration in the future.

REFERENCES

[1] Liliana Ardissono, Anna Goy, Giovanna Petrone, Marino Segnan, and Pietro Torasso. 2003. Intrigue: personalized recommendation of tourist attractions for desktop and hand held devices. *Applied artificial intelligence* 17, 8-9 (2003), 687–714.

[2] Kenneth J Arrow. 1950. A difficulty in the concept of social welfare. *Journal of political economy* 58, 4 (1950), 328–346.

[3] Shlomo Berkovsky and Jill Freyne. 2010. Group-based recipe recommendations: analysis of data aggregation strategies. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 111–118.

[4] Robin Burke. 2002. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction* 12, 4 (2002), 331–370.

[5] Lucas Augusto Montalvão Costa Carvalho and Hendrik Teixeira Macedo. 2013. Users' satisfaction in recommendation systems for groups: an approach based on noncooperative games. In *Proceedings of the 22nd International Conference on World Wide Web*. 951–958.

[6] Yann Chevaleyre, Ulle Endriss, Jérôme Lang, and Nicolas Maudet. 2007. A short introduction to computational social choice. In *International Conference on Current Trends in Theory and Practice of Computer Science*. Springer, 51–69.

[7] Fred D Davis. 1985. *A technology acceptance model for empirically testing new end-user information systems: Theory and results.* Ph.D. Dissertation. Massachusetts Institute of Technology.

[8] Amra Delic, Francesco Ricci, and Julia Neidhardt. 2019. Preference Networks and Non-Linear Preferences in Group Recommendations. In *IEEE/WIC/ACM International Conference on Web Intelligence*. ACM, 403–407.

[9] F Faul, E Erdfelder, AG Lang, and A Buchner. [n.d.]. A flexible statistical power analysis program for the social, behavioral and biomedical sciences. *Behavior Research Methods* ([n. d.]).

[10] Alexander Felfernig, Ludovico Boratto, Martin Stettinger, and Marko Tkalčič. 2018. Algorithms for Group Recommendation. In *Group Recommender Systems*. Springer, 27–58.

[11] Alexander Felfernig, Ludovico Boratto, Martin Stettinger, and Marko Tkalčič. 2018. Explanations for Groups. In *Group Recommender Systems*. Springer, 105–126.

[12] Shanshan Feng and Jian Cao. 2017. Improving group recommendations via detecting comprehensive correlative information. *Multimedia Tools and Applications* 76, 1 (2017), 1355–1377.

[13] Lidia Fotia, Fabrizio Messina, Domenico Rosaci, and Giuseppe ML Sarné. 2017. Using local trust for forming cohesive social structures in virtual communities. *Comput. J.* 60, 11 (2017), 1717–1727.

[14] Junpeng Guo, Lihua Sun, Wenhua Li, and Ting Yu. 2018. Applying uncertainty theory to group recommender systems taking account of experts preferences. *Multimedia Tools and Applications* (2018), 1–18.

[15] Daniel Herzog and Wolfgang Wörndl. 2019. A User Study on Groups Interacting with Tourist Trip Recommender Systems in Public Spaces. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, 130–138.

[16] Joseph A Konstan, Bradley N Miller, David Maltz, Jonathan L Herlocker, Lee R Gordon, and John Riedl. 1997. GroupLens: applying collaborative filtering to Usenet news. *Commun. ACM* 40, 3 (1997), 77–87.

[17] Judith Masthoff. 2004. Group modeling: Selecting a sequence of television items to suit a group of viewers. In *Personalized digital television*. Springer, 93–141.

[18] Judith Masthoff. 2011. Group recommender systems: Combining individual models. In *Recommender systems handbook*. Springer, 677–702.

[19] Judith Masthoff. 2015. Group recommender systems: aggregation, satisfaction and group attributes. In *recommender systems handbook*. Springer, 743–776.

[20] Judith Masthoff and Albert Gatt. 2006. In pursuit of satisfaction and the prevention of embarrassment: affective state in group recommender systems. *User Modeling and User-Adapted Interaction* 16, 3-4 (2006), 281–319.

[21] Kevin McCarthy, Maria Salamó, Lorcan Coyle, Lorraine McGinty, Barry Smyth, and Paddy Nixon. 2006. Cats: A synchronous approach to collaborative group recommendation. In *Florida Artificial Intelligence Research Society Conference (FLAIRS)*. 86–91.

[22] Shabnam Najafian and Nava Tintarev. 2018. Generating Consensus Explanations for Group Recommendations: an exploratory study. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, 245–250.

[23] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. 157–164.

[24] Lara Quijano-Sánchez, Belén Díaz-Agudo, and Juan A Recio-García. 2014. Development of a group recommender application in a social network. *Knowledge-Based Systems* 71 (2014), 72–85.

[25] Christophe Senot, Dimitre Kostadinov, Makram Bouzid, Jérôme Picault, Armen Aghasaryan, and Cédric Bernier. 2010. Analysis of strategies for building group profiles. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 40–51.

[26] Young-Duk Seo, Young-Gab Kim, Euijong Lee, Kwang-Soo Seol, and Doo-Kwon Baik. 2018. An enhanced aggregation method considering deviations for a group recommendation. *Expert Systems with Applications* 93 (2018), 299–312.

[27] Barry Smyth and Paul McClave. 2001. Similarity vs. diversity. In *International conference on case-based reasoning*. Springer, 347–361.

[28] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*. 22–32.