Integrating Collaboration and Leadership in Conversational Group Recommender Systems

DAVID CONTRERAS, Facultad de Ingeniería y Arquitectura, Universidad Arturo Prat, Chile
MARIA SALAMÓ, Facultat de Matemàtiques i Informàtica, Universitat de Barcelona, Spain
LUDOVICO BORATTO, Department of Mathematics and Computer Science, University of Cagliari, Italy

Recent observational studies highlight the importance of considering the interactions between users in the group recommendation process, but to date their integration has been marginal. In this article, we propose a collaborative model based on the social interactions that take place in a web-based conversational group recommender system. The collaborative model allows the group recommender to implicitly infer the different roles within the group, namely, collaborative and leader user(s). Moreover, it serves as the basis of several novel collaboration-based consensus strategies that integrate both individual and social interactions in the group recommendation process. A live-user evaluation confirms that our approach accurately identifies the collaborative and leader users in a group and produces more effective recommendations.

CCS Concepts: • **Information systems** → **Recommender systems**; **Personalization**; **Collaborative filtering**; **Social recommendation**; • **Human-centered computing** → **Collaborative filtering;**

Additional Key Words and Phrases: Group recommendation, interactions, collaboration, leadership, live-user evaluation

ACM Reference format:
David Contreras, Maria Salamó, and Ludovico Boratto. 2021. Integrating Collaboration and Leadership in Conversational Group Recommender Systems. *ACM Trans. Inf. Syst.* 39, 4, Article 41 (August 2021), 32 pages.
https://doi.org/10.1145/3462759

41

1 INTRODUCTION

Conversational recommender systems have been receiving more attention in the past few years [7, 10, 56, 60]. This is mainly due to the fact that the most common recommender systems, based on a one-shot interaction process, present limitations in domains where the preferences of the users cannot be reliably estimated from their past interactions. Enabling a conversational interaction process in a recommender system provides benefits, like improving the preference elicitation process [27].

Ludovico Boratto work partially done while at EURECAT - Centre Tecnològic de Catalunya.

This research was partially funded by Project No. 2017-SGR-341, MISMIS-LANGUAGE (Grant No. PGC2018-096212-B-C33) from the Spanish Ministry of Science and Innovation, NanoMoocs (Grant No. COMRDI18-1-0010) from ACCIO.

Authors' addresses: D. Contreras, Facultad de Ingeniería y Arquitectura, Universidad Arturo Prat, Avenida Arturo Prat, 2120, Iquique, Chile; email: david.contreras@unap.cl; M. Salamó, Facultat de Matemàtiques i Informàtica, Universitat de Barcelona, Gran Via de les Corts Catalanes, 585, Barcelona, Spain; email: maria.salamo@ub.edu; L. Boratto, Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, Cagliari, Italy, 09126; email: ludovico.boratto@acm.org.

1046-8188/2021/08-ART41 $15.00
https://doi.org/10.1145/3462759

Even though conversational recommendation is, in the vast majority of the cases, linked to single user recommendation, its connection with group settings is becoming prominent. Specifically, the recent survey by Jannach et al. highlighted the use of conversational recommenders to support group decision processes as an open research issue [27], and Delić et al. highlighted the need for group recommender systems to (i) support the decision-making process, (ii) identify individual characteristics inside a group, and (iii) account for choice satisfaction in groups [16].

Group recommender systems provide suggestions in contexts in which more than one person is involved in the recommendation process [11, 18, 44]. These systems aim to provide recommendations to the whole group, thus considering the preferences and the characteristics of more than one user. Most of the existing group recommender systems in the literature consider offline interactions between users and items, either explicit (e.g., via a rating) or implicit (e.g., the items they interacted with). Part of the literature, instead, has tried to go beyond user-item interactions, by considering concepts such as personality, emotion, and group dynamics [57]. In particular, several observational studies highlighted the importance of considering the interaction between the users in the final-decision making [12–15]. These studies also highlighted that *none of the existing strategies to model group preferences is able to infer the consensus that a real group would achieve through interactions.*

One recent approach in this direction has been proposed by Nguyen and Ricci [39], who considered the interactions of the users with the items in a chat-based group environment. However, the users did not really interact among themselves in a one-to-one interaction, but provided a feedback to the items that other users proposed to the whole group. The amount of feedback allows the system to infer who is the most active user in the group. An active user is a group member that highly interacts with items over the session. This interaction can happen by providing ratings or preferences over the items and is not necessarily a collaboration with another user within the group. While we acknowledge the authors for embracing interactions in the group recommendation process, the following aspects remain open: (i) instead of using solely the explicit feedback on the items evaluated by the other group members, it can also be important to take into account all latent forms of interaction that appear (otherwise, implicit feedback coming from interactions is lost); (ii) according to the mentioned approach, the most active users are the ones who are weighted more in the final recommendation; however, being active is not enough and a leader in a group needs to be recognized as such (hence, the response of the other users to her actions is a key aspect).

The following example shows the difference between considering as leader the most active user or the user whose items were appreciated the most.

> **Example**. *Consider a group of three users (u_1, u_2, u_3); their activities in a session are as follows. First, two users suggest products; u_1 suggests p_1 and u_2 suggests p_2. Next, u_1 views or rates p_2, and suggests another product p_3. Meanwhile, u_3 decides to view or rate p_2 and she also stores p_2 as one of her preferred products. Considering the activities of this session, u_1 is the most active user, since she has suggested p_1 and p_3, and has viewed p_2; indeed, u_1 has done 3 of the 6 activities of the group. User u_3 is the second most active. However, note that u_2 has suggested a product p_2 that has been viewed by the other group members (both u_1 and u_3) and it was also stored by u_3 as one of her preferred products. Therefore, it would be reasonable to consider u_2 as the leader of the group, since she has gained the attention of the group members and proposed a consensus product.*

To close the gap with respect to the open aspects we have previously highlighted, in this article, we analyze online users' interactions to support consensus agreement in a conversational group recommender system. Specifically, the interactions between the users are exploited by a novel collaboration model, which we integrate into consensus strategies, thus producing group recommendations that take into account both the individual preferences and the interactions. The foundation for this work is an online web-based conversational group recommender system, based on critiques, called gCOACH [9, 42]. A critique is an individual user's interaction to define a directional preference over one of the features, e.g., "*like this but cheaper.*" A conversational system like gCOACH allows us to have an environment in which users can (i) interact on a one-to-one basis thanks to the interaction modalities included in the web-based system (this allows us to collect implicit feedback on collaboration and leadership of the users in the group), and (ii) refine their choices by providing feedback and making suggestions to a specific group member; we consider the response to the item suggestions made by a user as a means to go beyond simple definitions of leadership that are based on who is more active. This allows us to overcome both the aforementioned open issues, by providing invaluable insights on group members' interactions. While gCOACH enables user interactions, our contribution in this article *includes them in group modeling and consensus strategies, thus reshaping the group recommendations based on group interactions*. To the best of our knowledge, no recommendation algorithm has previously considered the interaction between the users to integrate collaboration and leadership in an online group recommendation process. Specifically, in order for us to define a collaborative model from which we extract notions of collaboration and leadership, we consider two types of interactions, i.e., individual interactions to define their individual preferences, and social interactions to maintain awareness within the group. A social interaction is any interaction that generates awareness of a group member's decisions and allows this member to provide suggestions to others. We consider a user is collaborative when she is active in social interactions and, in addition, her interactions produce a reaction from those members she interacts with. We consider as leader the most collaborative user in a group.

More specifically, our contributions are summarized as follows:

- This is the first time that interactions (i.e., actions and reactions) between users during a session are exploited by a group recommender system;
- We model interactions between users and build a new collaboration model, from which we infer who are the collaborative users and who is a leader within the group;
- We present new collaboration-based consensus strategies, which consider both the individual and the collaboration models;
- We present a live-user evaluation to validate our proposal, which is a key tool to make an assessment on (i) the dynamics between users in the real-world and (ii) their impact in the group recommendation process, also considering the lack of live user experience research conducted in the context of group recommender systems.

The rest of the article is structured as follows. Section 2 presents related work. In Section 3, we present our problem definition. In Section 4, we provide the details of our approach to model collaboration and leadership and our novel consensus-based strategies. Section 5 presents the live-user evaluation and Section 6 its results. Finally, Section 7 concludes the article.

2 RELATED WORK

This section covers the main topics of this article, i.e., conversational group recommender systems that support feedback in the form of critiques, personality, social connections, and group dynamics in group-based algorithms. Readers can refer to the recent survey by Jannach et al. for conversational approaches that go beyond critiques [27]. Moreover, we refer to the recent book by Felfernig et al. [18] and survey by Dara et al. [11] for a more complete overview on the group recommendation research area.

Critique-based recommender systems. One of the main representatives of conversational recommenders are critique-based recommender systems, as stated by Chen and Pu in their survey [6]. This class of conversational systems is also referred to as critiquing-based recommendation in the literature. A critique-based recommender system supports feedback in the form of critiques. Critiques may be described in natural language or by using buttons in an interface. Mostly, researchers have been engaged in the design and development of interactive critiquing in graphical user interfaces, to better facilitate the process in such environments. The rationale behind the use of critiques is based on the idea that it is easier for a user to critique a product recommendation by saying "*like this but cheaper*" than to construct formal queries [4]. A significant amount of the research carried out on recommender systems has widely recognized the benefits of critique-based recommenders [6, 37, 47], since they have proven to be especially helpful for users with ill-defined needs and preferences. Critique-based group recommender systems [19] have been developed in different domains such as recommending vacations to groups of tourists [35] and recommending restaurants [23]. ***Collaborative Advisory Travel System (CATS)*** [34, 35] helps a group of users in planning a skiing vacation. *Where2Eat* [23] is a mobile application that recommends restaurants. *gCOACH* is a COllaborative Advisory CHannel for group recommendations [9] that supports online conversational group recommendations scenarios in a web-based environment. *Choicla* [55] supports different group decision scenarios; even if it is not based on critiquing, *Choicla* enables groups members to evaluate different item features. In *Hootle+* [1], preference elicitation and negotiation is done by enabling group members to accept or reject the proposed features and adjust their significance.

Personality and social relationships in group recommendation. Personality reflects "individual differences in emotional, interpersonal, experiential, attitudinal, and motivational styles" [36]. To obtain personality information and consider it in the recommendation process, questionnaires are used [57] and models such as the Five Factors (Big Five) and Thomas-Kilmann Conflict Resolution Style are usually considered. Sánchez et al. [50] consider the Thomas-Kilmann model to enrich the single-user ratings with the influence and conformity of a user when she is part of a group. Nguyen and Ricci [40] study the relation between the conformity of the users and the preferences to be used in the group model. Delic et al. [15] show that user personality is related to the satisfaction of the individual users with the group recommendations.

Social relationships can play a role when modeling group preferences. Gartrell et al. [22] discovered that the type of relationship between users can be used to decide the group modeling strategy. Sánchez et al. show that recommendation effectiveness can be improved by considering both personality and social trust [51] or by inferring notions of trust from Facebook [49]. Christensen and Schiaffino [8] improve group recommendation accuracy with a notion of social influence that includes social trust, social similarity, and social centrality. Barile et al. [2] consider tie strength, relationship type, and closeness of initial ratings.

Collaboration in conversational search. In the area of collaborative search, different studies have analyzed the effect of collaboration in information seeking [52], the strategies users adopt to maximize their knowledge gain in collaborative web search sessions [59], the impact of group size on collaborative search effectiveness [38], and the role of collaborators implied within a collaborative search session [53, 54]. In this article, our point of analysis is different; instead of analyzing the users' role in searching information, we focus on analyzing the influence of one or more group member (i.e., leader(s)) over the remaining members of the group. In other words, it is important to understand the social influences during interaction to make a group decision. In group recommendation scenarios, a unique final decision should be reached among members.

User interactions in group recommendation. The last aspect we consider are the interactions between users. Several observational studies [13–15] have analyzed how the interactions between users shape the decision-making process; results show that classic group modeling strategies are not able to model the consensus the group achieves by interacting, and that personality of the users drives the group decisions. Another observational study [12] analyzed the impact of social centrality in the group decision-making process, which does not have an impact on groups characterized by notions of equality; however, when groups are loosely tied, socially-central group members are happier with the final decision. Other approaches have focused on the interaction between users to decide the best item for a group [58], thanks to deep learning applied to collaborative filtering. The last approach we consider, by Nguyen and Ricci [39], is the most similar to ours. This mobile chat-based group recommender system allows users to propose POIs, which can be up- and down-voted by the other group members. The preferences expressed by the users prior the chat and the feedback provided during it are used to model the user preferences in terms of tourism features. Group recommendations are computed by considering these models, weighted by interaction (the more a user interacted with a group, the higher is the importance of her preferences when computing the recommendations).

Contextualizing our contribution. Our system goes beyond the use of the conversational process to understand the user and provide recommendations, as classic approaches do [6, 37, 47]. In our contribution, instead, the collaboration process is used to extract knowledge on the users, in terms of their collaborative and leadership characteristics. With respect to the existing works on personality, we do not add notions of it in the group model, but we analyze a-posteriori if personality has an impact in the group decision making. Regarding the prior social relationships between the users (e.g., friendship), we do not explicitly consider them, but just observe how users interact in the group and implicitly infer them. The studies related to group dynamics are those that overlap the most with our study; with respect to the observational studies, we go beyond simple observation of the interactions, by exploiting them to build a collaborative model that is used to infer the leader of the group. The approach presented in Reference [58] employs a history log, i.e., no live user interaction between group members is possible. With respect to the study presented in Reference [39], the main difference is on who is recognized as the most relevant user; indeed, the previous work considered the most active user, while we consider also the response that the other group members had on the actions she made.

3 PROBLEM DEFINITION

We consider a group recommender system designed to *provide suggestions in scenarios where a group of users is engaged* [18, 26, 44]. In these scenarios, typically a group of people are intending to participate in a group activity [41] (e.g., to suggest a destination to a group of people who want to have a trip together).

3.1 Background on Conversational Group Recommendation

Before presenting the problem we tackle in this study, we introduce some background on a classic conversational group recommendation framework.

We consider a set of users, $\mathcal{U}$, whose goal is to search for a unique item i (extracted from the large set of items $\mathcal{I}$) that maximizes the satisfaction of the preferences of the whole group.

Users are connected through a synchronous recommendation session. During the session, a user gets a suggestion of an item and provides feedback to express their preferences on the recommended item. This feedback is used by the group recommendation algorithm to suggest the user a new item that satisfies her preferences. The process is repeated until the user is satisfied with an item, which she selects as her choice; this choice denotes the end of the session for that user. Nevertheless, the group recommendation process continues until each group member selects one item that satisfies her preferences. Clearly, the items selected by the individual users can be different (in other words, not all the users select the same item as their favorite). For that reason, group recommender systems introduce an additional phase, which involves reaching a consensus among members. This allows the system to find an agreement on what item should be proposed to the group. More formally, the setting can be defined as follows: let $\mathcal{U} = \{u_1, \ldots, u_n\}$ be a set of users, where u_i represents the ith user and n the number of users in the group, $|\mathcal{U}| = n$. Let $\mathcal{I} = \{I_1, \ldots, I_m\}$ be a set of items for recommendation, where I_j is the jth item. Depending on the recommendation algorithm, $\mathcal{I}$ will be the whole set of items or a subset of them ($|\mathcal{I}| \leq m$).

The set of individual user models is defined as $\mathcal{IM} = \{IM^{u_1}, \ldots, IM^{u_n}\}$, where IM^{u_i} represents the individual model of the ith user. We consider that each user u_i has an individual user model $IM^{u_i} = \{P_1, \ldots, P_{r_{u_i}}\}$ that represents her individual preferences, from her initial preference P_1 to her last preference $P_{r_{u_i}}$. Note that the subindex r_{u_i} may be different for each user in $\mathcal{U}$, because each user defines her own set of preferences. Thus, R is the total number of preferences for the group, which is computed as $R = |IM^{u_1}| + \cdots + |IM^{u_n}| = r_{u_1} + \cdots + r_{u_n}$.

To understand if an item $I_j \in \mathcal{I}$ is relevant for a user $u_i \in \mathcal{U}$, we employ a satisfaction measure, defined in Equation (1), which measures how many of the user's preferences are satisfied by a product:

$$\delta(I_j, IM^{u_i}) = \sum_{s=1}^{r_{u_i}} w_s \cdot \text{Satisfies}(I_j, P_s). \tag{1}$$

Specifically, $\delta(I_j, IM^{u_i})$ shows, for a particular item I_j and a user u_i, the number of preferences from her individual user model IM^{u_i} that product I_j satisfies. The parameter w_s is a weighting factor that can be used to tune preferences. The $Satisfies(I_j, P_s)$ function depends on the type of feedback the recommender supports. For example, (1) in a conversational content-based recommender system that uses *critiquing-based* feedback, this measure computes whether a critique P_s is satisfied by item I_j, while (2) in a collaborative filtering recommender that uses *rating-based* feedback, this measure directly returns the rating value. A rating, P_s, is a numerical score provided by the user that measures how much user u_i likes an item I_j. In this particular case, the preference provided also denotes the satisfaction of the user.

The goal of the group recommendation algorithm is to reach consensus in the group. The consensus item recommendation to a group is generated according to Equation (2), whose focus is on maximizing the preferences of all group members:

$$\text{consensus}(\mathcal{I}, \mathcal{IM}) = \arg\max_{I_j \in \mathcal{I}}(\text{strategy}(I_j, \mathcal{IM})), \tag{2}$$

where *strategy* refers to the name of the strategy used to aggregate/combine the individual recommendations.

3.2 Problem Formulation

As we have shown in our background, classic conversational group recommenders only deal with what we defined as *individual interactions* (i.e., those interactions where the user provides her own preferences over the items). However, the whole motivation behind this work is to be able to capture also *social interactions*. For this reason, the goal of this article is dual. First, we need to *capture the interaction between the users*. This is done by defining a collaborative model (from now on $\mathcal{CM}$) that stores different types of interaction. The second and final goal of this article is to use $\mathcal{CM}$ to score each user in the group in terms of her role as *collaborator* or *leader*. This allows us to define the influence of each user in the group. Formally, we enrich Equation (2) by adding the collaborative model $\mathcal{CM}$. This collaborative model enables redefining the consensus phase, see Equation (3), and to propose new collaborative-based consensus strategies, see Section 4.4:

$$\text{consensus}(\mathcal{I}, \mathcal{IM}, \mathcal{CM}) = \arg\max_{I_j \in \mathcal{I}}(\text{strategy}(I_j, \mathcal{IM}, \mathcal{CM})), \tag{3}$$

where *strategy* refers to the name of the strategy used to aggregate/combine the individual recommendations, $\mathcal{IM}$ is the set of individual models, and $\mathcal{CM}$ is the new collaborative model.

3.3 Research Questions

With the definition of our collaboration model $\mathcal{CM}$, and the assessment of who are the collaborative and leader users in the group, we aim to address the following research questions:

- **RQ1.** Can we infer efficiently[1] who is a leader or a collaborator in a group, by analyzing social users' interactions?
- **RQ2.** Is a leader characterized by a clear psychological conflict style mode?
- **RQ3.** Can a leader improve the effectiveness of group recommender systems in a consensus-based decision-making process?

3.4 Hypotheses

Before describing the collaborative model, we now list the specific hypotheses we investigate in this work, based on the goal we would like to address, i.e., analyzing social influence in conversational group recommendation scenarios.

- **H1.** By monitoring social users' interactions it is possible to build a collaborative model and infer notions of *collaboration* and *leadership* in a group.
- **H2.** The detection of leader(s) in a group improves the effectiveness of consensus strategies in conversational group recommender systems.

Considering the relationship between hypotheses and research questions, H1 is related to RQ1 and RQ2. In RQ1, we monitor social user's interactions to characterize who is a leader or a collaborator. In RQ2, we analyze if users are collaborators taking into account the psychological perspective. In particular, we analyze if a user is collaborator considering their individual behavior when responding to a conflict situation, such as reaching consensus agreement in a group recommender system. However, H2 is related to RQ3. We analyze the impact of considering leader satisfaction, instead of the whole group satisfaction, in the consensus-based decision-making process.

4 COLLABORATION-BASED GROUP RECOMMENDATION

In this section, we present our approach to reach a collaboration-based group recommendation. Section 4.1 presents how we model social interactions and Section 4.2 describes how we build the collaborative model. The process for identifying collaboration and leadership from the collaborative model of interaction is described in Section 4.3. Finally, we present different collaboration-based consensus strategies in Section 4.4, which integrate the collaborative model to provide group recommendations.

4.1 Modeling Interaction

We assume that users participate in a group activity through an interface that enables different social interactions. In this article, we have used a framework (see Section 5.3) that is synchronous

[1]With *efficiently*, we mean in a quick and accurate way. Our measure of accuracy is defined in Equation (19).

and includes different social interaction modalities to maintain awareness and to suggest items among members.

As described in Section 3.1, we have a set of users, $\mathcal{U}$. For each user u_i, we define a user model, $UM^{u_i} =< IM^{u_i}, CM^{u_i} >$ that contains two components: the **individual model (*IM*)** focused on storing the individual interactions, and the **collaborative model (*CM*)** that concentrates on the social interactions.

The set of individual user models is described in Section 3.1. Recall that an individual user model IM^{u_i} stores the individual preferences of user u_i during the session. The set of collaborative user models will be defined as $\mathcal{CM} = \{CM^{u_1}, \ldots, CM^{u_n}\}$, where CM^{u_i} represents the collaborative user model of the ith user. Each user u_i has a collaborative user model that represents her social interactions to the remaining group members and how these interactions have been viewed and accepted by each individual user.

Since the social interactions depend on the type of interface, we have focused on the social interactions that are the most likely to appear at any group recommendation interface. In this context, conversational interfaces include mechanisms for sharing information among group members (e.g., to recommend or suggest product(s) from one member to another, or to recommend from one member to the whole group). Once a member suggest a product to one or more members, it may generate the reaction of the members when looking at the suggestion. That is, there are social interactions that happen as a result of a previous one. For example, to open or rate a product suggested by someone; and to save or choose the product suggested to the user by another group member. These forms of interaction can support group decision-making processes via conversational interfaces. While it is true that many other forms of social interactions may exist, their analysis is out of the scope of this article. Hence, this work considers the main types of social interactions, namely, to suggest products, to view suggestions, and to store them if they are of interest to the user. Formally, a collaborative user model for a user u_i is described as a triplet:

$$CM^{u_i} = < SU, VP, SP >, \tag{4}$$

where SU represents the number of product suggestions made by the user u_i to any group member u_j, VP is the number of viewed products (i.e., those items suggested by u_i that has been viewed by the target member u_j), and SP is the number of stored products (i.e., items suggested by u_i that have been added to her preferred list of items by the user u_j). We denote the access of any of these three components as $CM^{u_i}_{SU}$, $CM^{u_i}_{VP}$, and $CM^{u_i}_{SP}$. Note that the initial value for the three components is set to 0 at the beginning of the group recommendation activity.

4.2 Building the Collaborative Model

The group recommendation process has two phases. The first phase is devoted to the users' interaction, where the users navigate through the product space and interact between them to locate their desired products. The second phase starts once each user has finished their session by selecting one or more preferred items and it is devoted to reach consensus between the group members.

At the first phase, users receive recommendations based on their individual models $\mathcal{IM}$ (formally described in Section 3.1) and a group user model $GUM = \{G_1, \ldots, G_n\}$, where G_i is a set of the preferences made by user u_i. $\mathcal{IM}$ and GUM are empty at the beginning of the session. In both models, $\mathcal{IM}$ and GUM, prior to adding a new preference, all existing preferences that are inconsistent with it are removed. Note that the recommendation generation process at this phase is based on the individual interaction and the social interactions are not considered at all. Inconsistencies happen when the preference contradicts or refines a previous stored preference in the IM^{u_i} or GUM models (i.e., u_i's individual model or the group user model). There are two types of inconsistencies: contradictions and refinements. The former happens when the latest preference contradicts a previous one in the model. In that case, previous preferences are removed from the model. The last preference is maintained, because the recommender assumes that users may refine their requirements over time, since they are learning about the product space during the conversational recommendation process. The latter inconsistency is a refinement and it happens when there is an improvement over one of the preferences. For example, considering a model that contains preferences defined as critiques and the individual user model contains a critique *(price; <; \$300)*. Then, if the user performs a critique *(price; <; \$200)* over the current recommendation. In this case, the critique *(price; <; \$300)* stored in the individual model will be refined to *(price; <; \$200)*.

Both the individual ($\mathcal{IM}$) and the collaborative ($\mathcal{CM}$) models are built during the first phase, when the individual and social interactions happen. To build up the collaborative model ($\mathcal{CM}$) is necessary to store for each member of the group whether her social interactions come from suggestions made by one user to another. Formally, we define this notion as $SUG = \{SUG^{u_1}, \ldots, SUG^{u_n}\}$, where each SUG^{u_i} represents the suggestions that user u_i has received. Let $SUG^{u_i} = \{I_1, \ldots, I_g\}$ be the set of items that u_i has received either from others or by the recommendation algorithm. Note that a user is not allowed to suggest items to herself.

We update CM every time a social interaction happens in the group recommender system. Formally, every time a user u_i sends a product suggestion I_k to a user u_j, we apply Equation (5):

$$SUG^{u_j} = SUG^{u_j} \cup I_k^{u_i}. \tag{5}$$

That is, we add the item $I_k^{u_i}$ to the suggestions set SUG^{u_j} of user u_j. A product suggestion maintains awareness within the group and it enables collaboration in the group. Accordingly, for every product suggested to another user, $I_k^{u_i} \in SUG^{u_j}$, we also update the number of suggestions made by user u_i:

$$CM_{SU}^{u_i} = CM_{SU}^{u_i} + 1. \tag{6}$$

The larger the $CM_{SU}^{u_i}$, the more collaborative user u_i is. Once suggested an item, three different situations may happen. First, the item $I_k^{u_i}$ is not viewed and not stored in the set of preselected items (from now on called the stack) by user u_j. That means I_k is not of interest to u_j. Second, the item $I_k^{u_i}$ attracts u_j's attention and she views the product. With this situation, user u_i is more than a collaborative user, since she has influenced over the individual interaction of user u_j. Accordingly, every time a user u_j views a suggestion I_k made by a user u_i, we apply Equation (7):

$$CM_{VP}^{u_i} = CM_{VP}^{u_i} + 1. \tag{7}$$

The larger the $CM_{VP}^{u_i}$, the more influential user u_i is. The last situation happens when user u_j views product $I_k^{u_i}$ and u_j decides to add it to the stack, because it is of her interest. Hence, the product suggested by u_i has been followed by u_j. Accordingly, when a user u_j stacks a suggestion I_k made by a user u_i, we apply Equation (8):

$$CM_{SP}^{u_i} = CM_{SP}^{u_i} + 1. \tag{8}$$

The larger the $CM_{SP}^{u_i}$, the more followed the user is.

Although the collaborative model is updated during the first phase of the recommendation process, the purpose of this article is to analyze whether social interactions influence positively in improving the final decision-making in group recommendation activities. For this reason, we analyze social influence at the second phase of the group recommendation process, where the focus is on reaching consensus. In our future work, we will address this analysis during the first phase of the process.

At the second phase, the common recommendation generation approach to reach consensus is defined in Equation (2), which is based solely on the individual interactions. In this article, we redefine this equation to integrate social interactions in addition to individual interactions ($\mathcal{IM}$

and $\mathcal{CM}$), as defined in Equation (3). Instead of using the whole set of items $\mathcal{I}$, as we are reaching consensus, we only use the set of items selected at the end of the first phase by all the users. This set is formally defined as $STACK = \{I_1, \ldots, I_s\}$, which stores the preselected items of the group members, where I_s is the sth item. The items included in $STACK$ and SUG contain the information of the product and the identifier of the user that has introduced it. An item j that comes from a particular user i is denoted as $I_j^{u_i}$. The stack is used to reach consensus among users and, as defined in Equation (9), the final recommendation generation will be one of the items stored in the stack:

$$\text{consensus}(STACK, \mathcal{IM}, \mathcal{CM}) = \arg \max_{I_j \in STACK}(\text{strategy}(I_j, \mathcal{IM}, \mathcal{CM})). \tag{9}$$

4.3 Collaboration and Leadership

Thanks to $\mathcal{CM}$, we can extract a collaborative score for each user:

$$\text{score}_{u_i}(\mathcal{CM}) = CM_{SU}^{u_i} \cdot w_{su} + CM_{VP}^{u_i} \cdot w_{vp} + CM_{SP}^{u_i} \cdot w_{sp}, \tag{10}$$

where w_{su}, w_{vp}, and w_{sp} weigh each component differently. The particular weight values used for our experiments are detailed in Section 5.

As defined by Kapetanios [28], a collaboration is a *true synergy among diverse participants in creating solutions or strategies through the synergistic interactions of a group of people.* Our definition of collaborator is in this vein, since we consider a collaborator is a member that has social interaction with other members to reach consensus.

From the score defined in Equation (10), we define the leader as the most collaborative user in the group (i.e., as the user u_i with the maximum collaborative score), see Equation (11):

$$\text{leader}(CM) = \arg \max_{u_i \in U}(score_{u_i}(CM)). \tag{11}$$

Our definition of leader, see Equation (11), is rooted in studies of social influence in groups [33], in the social psychology field. However, many definition and approximations to the concept of influence and leadership exist in social psychology [20, 33]. In their survey, Hogg [25], has highlighted that leaders are *agents of influence* and, when people are influenced, it is often because of *effective leadership*. However, Chemers [5] defined leadership as "*a process of social influence through which an individual enlists and mobilizes the aid of others in the attainment of a collective goal.*"

We conceived our definition of leader by taking into account the concepts described by Hogg [25] and Chemers [5]. We consider that a leader is a member that influences other people to pursue a common goal. Thus, any member of the group can exhibit some level of leadership at any time, since all of them may provide suggestions (as captured by the $CM_{SU}^{u_i}$ component of the collaborative model). However, a leader must not only be an active user by suggesting products to the group, she also needs to influence the remaining members. This influence is achieved when the group members view the products suggested (i.e., $CM_{VP}^{u_i}$) and stack them (i.e., $CM_{SP}^{u_i}$). Accordingly, our definition considers that the leader is the most collaborative, influential, and followed member of the group.

4.4 Collaboration-based Consensus Strategies

To date, little work has been addressed to include social influence within group recommender systems (as we presented in Section 2). Here, we integrate collaboration into well-known consensus strategies as well as we define two new proposals in the field, based on the collaboration and leader definition detailed in Equation (11). To the best of our knowledge, this is the first time online social interactions has been modeled and integrated in consensus strategies.

Next, we define different strategies for reaching a consensus in a group recommender system. The consensus-recommended product(s) to a group is generated according to Equation (9), where *strategy* refers to the name of the consensus proposals (i.e., *mean*, *completeness*, *multiplicative*) described below. Recall that the consensus happens at the second stage and it is based on the products that belong to the stack, $STACK \subseteq \mathcal{I}$.

4.4.1 Collaborative Mean. The *collaborative mean* satisfaction, $mean(I_j, IM, CM)$, of the group for a product, I_j, is defined as the sum of each member's individual satisfaction according to her individual preferences, IM^{u_i}, averaged by the degree of collaboration of every user, which is extracted from the collaborative user model CM^{u_i} (recall that n denotes the number of users in a group):

$$\text{mean}(I_j, IM, CM) = \frac{1}{n} \sum_{i=1}^{n} \delta(I_j, IM^{u_i}) \cdot \text{score}_{u_i}(CM). \tag{12}$$

The satisfaction measure δ is defined in Equation (1) and measures how much u_i likes the product I_j based on their set of preferences stored in IM^{u_i}.

4.4.2 Collaborative Completeness. Equation (13) defines the *collaborative completeness* consensus-based proposal. The objective of the collaborative completeness strategy is to favor high scores while penalizing big differences between members. Additionally, the strategy considers the collaborative score of each user, $score_{u_i}(CM)$:

$$\text{completeness}(I_j, IM, CM) = \frac{\sum_{i=1}^{n} \; w_i \cdot \sqrt{\delta(I_j, IM^{u_i})}}{\sum_{i=1}^{n} \sqrt{r_{u_i}}} \cdot \text{score}_{u_i}(CM), \tag{13}$$

where w_i is a weighting factor such that $\sum_{i=1}^{n} w_i = 1$ and $w_i \geq 0$. Recall that r_{u_i} is the total number of preferences of user u_i, as defined in Section 3.1. The completeness of an item, I_j, is computed in terms of members' satisfaction, $\delta(I_j, IM^{u_i})$. Moreover, the parameter w_i is a weighting factor used to tune the completeness—i.e., to focus more on one of the member's preferences (in our case, all group members take the same value for w_i). The sum of $\sqrt{r_{u_i}}$ factors is for normalizing the values in the range [0,1].

4.4.3 Collaborative Multiplicative. This strategy multiplies the satisfaction of the individual users and it is weighted according to the collaborative score, see Equation(14). With this strategy, it might happen that a member with unique tastes always loses out, because their opinion happens to be a minority preference. However, if the other ones have a low collaborative score, it may increase their influence:

$$\text{multiplicative}(I_j, IM, CM) = \prod_{i=1}^{k} (\delta(I_j, IM^{u_i})) \cdot \text{score}_{u_i}(CM). \tag{14}$$

4.4.4 Leader Satisfaction. Previous strategies (i.e., *mean*, *completeness*, and *multiplicative*) assume the consensus based on the satisfaction and the collaboration made by all the group members. Instead of analyzing the collaboration of the whole group, we propose two strategies based on the satisfaction of the leader, who has been the most collaborative and influential within the group.

In fact, the leader is chosen because she has been providing the group with suggestions, of which some or all have been viewed and stacked by the members of the group. We consider that the leader can be seen as an influencer for the group. For this reason, we propose two consensus-based strategies that only consider the preferences of the leader: *maximize leader satisfaction* and *leader selection.*

Both strategies look for maximizing the leader preferences, as denoted in Equation (15), which uniquely uses the individual model of the leader:

$$\text{leader\_satisfaction}(I_j, IM, CM) = \ \delta(I_j, IM^{u_l}) \text{ for a } u_l \text{ whose } \text{score}_{u_l}(CM) = \text{leader}(CM), \quad (15)$$

where u_l is the user ($u_l \in U$) whose collaborative score, obtained from $score_{u_l}(CM)$, corresponds to the leader score, computed using $leader(CM)$. In other words, the group member that has the highest collaborative score is considered to be the leader, as defined in Equation (11). The main difference between the two strategies is that they use a different subset of the products stored in the $STACK$.

Max leader satisfaction. In the *maximize leader satisfaction* strategy, the goal is to focus on the leader preferences but considering the whole set of products chosen by the group and stored in the stack. This strategy selects, among the products stored in the $STACK$, the one that maximizes the satisfaction of the leader's preferences, as denoted in Equation (15). The recommendation generation is based on Equation (9), that can be rewritten based on the *Max leader satisfaction* as depicted in Equation (16). This strategy generates a recommendation that maximizes the preferences of the leader although the selected product may be a product stored by another member of the group and not necessarily by the leader. Note that the leader might not have stacked a product of her interest, because it had already been added in the stack by another user. Hence, even though the *maximize leader satisfaction* strategy only depends on the leader, the final products might come from *any* user in the group:

$$\text{consensus}(STACK, \mathcal{IM}, \mathcal{CM}) = \arg \max_{I_j \in STACK} (\text{leader\_satisfaction}(I_j, \mathcal{IM}, \mathcal{CM})). \quad (16)$$

Leader selection. In the *leader selection* strategy, the idea is to consider both the preferences and the items chosen by the leader. Thus, guaranteeing the generation of a recommendation that is of interest for the leader. To this purpose, it is necessary to extract from the stack the subset of items chosen by the leader. The *leaderSet* is formally described in Equation (17):

$$\text{leaderSel} = \{I_j | I_j^{u_i} \in STACK \text{ and } u_i = u_l\}, \quad (17)$$

where $I_j^{u_i}$ is the item I_j that was added by user u_i to the stack. As a result, the *leaderSet* contains all items that the leader has added to the stack.

The *leader selection* strategy generates recommendations by considering only those products stored by the leader and, hence, the remaining ones will not be considered for consensus (see Equation (18)):

$$\text{consensus}(STACK, \mathcal{IM}, \mathcal{CM}) = \arg \max_{I_j \in \text{leaderSet}} (\text{leader\_satisfaction}(I_j, \mathcal{IM}, \mathcal{CM})). \quad (18)$$

The *leader selection* strategy selects among the products stacked by the leader, the one that maximizes the satisfaction of the leader's preferences, as denoted in Equation (18). As a result, the final product chosen for the group will be one of the selections made by the leader. Note that the *leader selection* strategy wields a lot of influence on the preferences made and products stacked by the leader. The remaining members of the group have no influence at all over the generated recommendation. The key idea of this restricted strategy is to analyze how users feel with a product exclusively chosen by the leader.

5 LIVE-USER EVALUATION

This section describes in depth our live-user study. Section 5.1 details the participants of our experiments. Section 5.2 indicates the domain and the parameters of the proposals. In Section 5.3, we introduce the group recommendation framework used in our experiments. Section 5.4 describes our online evaluation methodology. Finally, Section 5.5 details the metrics and tests applied in our evaluation.

5.1 Subjects

We recruited 68 participants, all of them students at the faculty of Engineering and Architecture at the University Arturo Prat (Chile). These participants were joined in 17 groups[2] of 4 participants for each test. The characteristics of the groups are described in Table 1. We requested participants to detail their age, the degree of relationship they have with the other group members (using a 5-point Likert scale where 1 indicates "no relationship" and 5 is "close friends"), the studies they have conducted (where 1 is a Pregraduate student, 2 is a Post-graduate, 3 is a Master student, 4 is a Ph.D. student, and 5 is a Ph.D.), how much they like skiing (using a 5-point Likert scale where 1 indicates "I hate SKI" and 5 is "I strongly love SKI."), and the level of ski that they have (where 1 indicates "no level" and 5 is "Master of SKI" in a 5-point Likert scale). As shown in Table 1, our participants are young people, because we recruited students at the university and their studies ranged from pre-graduate and post-graduate. Moreover, most of them like skiing and their level of ski is medium. The most important characteristic for our live-user study is the relationship column, which shows how much the members of the groups know each other. Since the participants are students at the same university, all of them have seen each other at the university. However, this does not mean that they are friends and know the preferences of each other. The relationship between the participants is the aspect that we have focused the most in the formation of the groups, ranging from groups whose members hardly know each other (with an average value of 2 in a 5-point scale) to those whose relationship is high (4 or above).

5.2 Domain and Setup

We have chosen the skiing domain in our live-user study, because travel vacations has been the most used scenario in conversational group recommender systems, such as References [26, 35, 46]. Moreover, the skiing domain has been previously also considered in other non-conversational group recommendation settings [34, 48]. Besides its popularity, a group vacation also represents a nice benchmark where people strongly interact in the shaping of their travel. Hence, our platform could be exploited to facilitate these interactions and evaluate our conversational approach. Concretely, the data set contains 153 European skiing holidays described by 41 features related to the resort (e.g., country or transfer time) and the accommodation (e.g., rating, price, or restaurant facilities).

In our live-user study, we setup the parameters of Equation (10) as follows: $w_{su} = 0.2$, $w_{vp} = 0.3$, and $w_{sp} = 0.5$ (the tuning was done on experiments conducted before this study and considering the framework presented in Reference [42]). Moreover, the initial value for the three components $CM^{u_i}_{SU}$, $CM^{u_i}_{VP}$, and $CM^{u_i}_{SP}$ of the collaborative user model is setup to 0. In addition, the w_i parameter in Equation (13) is setup to 0.25 for each member of the group.

5.3 Group Recommendation Framework

This section presents the conversational group recommendation framework [9, 42] used to collect information about the collaboration between users. The framework is called gCOACH (COllaborative Advisory CHannel for group recommendation). The framework supports online group recommendation scenarios and allows several users to participate in a group activity that involves searching a product for the whole group (see Figure 1).

[2]The live-user study was designed with 20 groups and a total of 80 users. Unfortunately, due to COVID-19 confinement and social restrictions, we were only able to reach 17 groups.

Table 1. Description of the Groups Involved in the Live-user Study

Group	Age	Relationship	Studies	Like Ski	Level Ski
1	20.25	2	2	3.5	3
2	26.25	2	2	2.5	2
3	22.5	2	1	2.75	2.25
4	23	2	1	2.25	2
5	21	2	1	2.75	2.25
6	20.75	2.25	1	2.75	2.25
7	21.75	3.25	3.5	3.5	3.5
8	20.5	4.25	1.25	3	1.25
9	32.25	2	2	2.75	2.25
10	21.25	2	1	2.75	2.5
11	20.75	3	1.5	3.25	3.25
12	21.25	3.75	1.75	3.5	3.75
13	22	4	1	3.25	3.25
14	20.75	3.25	1	3	2.5
15	20.75	3	1.5	2.5	2.25
16	21.25	3.75	1	3.25	2.5
17	22.25	4	1	2.75	2.5
Avg.	**22.25**	**2.85**	**1.44**	**2.94**	**2.54**

The characteristics of the users (Age, Relationship, Studies, Like Sky, and Level Sky) have been averaged among the four members of the group.

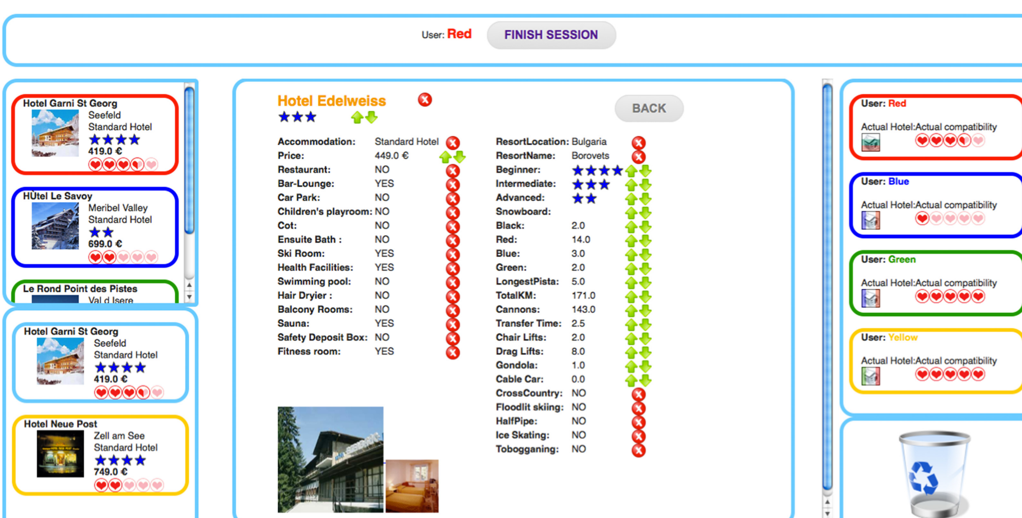


Fig. 1. The main interface of gCOACH.

The conversational nature of gCOACH guides users through a product space in pursuit of suitable products using a cyclical recommendation process (i.e., a dialogue), alternatively making suggestions and eliciting user feedback, to refine their needs and preferences, based on recent recommendations. In addition, gCOACH aims at being an on-line framework that facilitates group interaction and communication among members. gCOACH not only elicits users' preferences in the form of critiques during the session, in an environment that is domain-independent but also provides multiple interaction modalities and interface components to keep awareness of member's decisions and to provide suggestions among them. Note that social interaction modalities are used to maintain a conversation among group members. At the moment, gCOACH does not include a chat for enabling users to maintain a natural language conversation but it can be easily included in the same way we did in a conversational recommender system for single users [24].

Its architecture is a web-based environment, developed on a client-server model for enabling the interaction of the users from anywhere, details are described in Reference [9]. The architecture is divided into three main layers: a client, a server, and a recommender. Since we are interested in how users interact between them to infer knowledge for generating recommendations, we detail the individual and social interaction modalities (summarized in Table 2) in the client layer.

The user interface enables different individual or social interaction modalities and functionalities. Figure 1 depicts the interface shown to each user in the browser. The example interface shown is focused on a skiing package domain. First, note that each of the users is represented by a color (look at the top of the figure). In this example, the user is the red one. This color is used in the interface to denote to whom each element in the interface belongs. As depicted in Figure 2, the interface screen is divided into several areas: individual interaction, awareness, suggestions, stack, and waste basket, which are now presented in detail.

The area denoted with number 1 in Figure 2 is devoted to *individual interaction*, which describes a product in terms of its features and the value of each. Additionally, each one of the features contains one or two buttons for performing critiques. The user is able to make a critique, which allows her to express a preference over a feature in line with their personal requirements (e.g., *cheaper* or *higher star rating for hotel*, etc.). This feedback enriches the individual preference model and the group preference model. Next, the recommendation algorithm uses this feedback about the user taste in conversational fashion, and it answers to that feedback by replacing the product displayed with a new recommendation that better matches with the preference expressed.

Furthermore, when individual users arrive to a particular product recommendation that satisfies their requirements and wish to draw it to the attention of the other group members, they can do this by performing a drag and drop action. With this action, a user adds the recommended product to the suggestion box of another user, placed in the awareness area (denoted with number 2 and explained below) or to a stack area (denoted with number 5, also explained below), which is a social interaction modality.

Area number 2 in Figure 2 *keeps awareness* among members of the group. In short, this area is used by the target user to know what products the other users are currently viewing (i.e., the target user is aware of what the other are browsing) and to recommend them items (i.e., the target user makes the others aware of what could be of possible interest for them). This area contains a set of color boxes, each one representing a group member. Each color box shows which product each user is currently viewing. Users can browse this product by performing a click on it. Besides, the color box contains a 0–5 heart score that represents how compatible is the product to the user that is currently looking at it. The ultimate goal of this heart score is to know the products the users are currently interested in. This area is also used by other users to make suggestions of the current recommendation (displayed in area 1) to a specific user by doing a drag and drop of the product into the target user box. This suggested product will appear in the suggestions box of the target user, displayed in the suggestions area (denoted with number 4, explained below). Note that the awareness area not only maintains awareness between members, it also enables collaboration between the group members.

The *suggestion box* area is depicted in Figure 2 with number 4. These suggestions may come from any of the group members or as a result of a *proactive* suggestion of the recommendation

Table 2. Overview of the Interaction Modalities and the Action that Occurs in gCOACH Interface to Facilitate Group Interaction and Communication

	Interaction modality	**Area**	**Action**
1	Individual	#1	Area #1 is the individual area. User perform critiques to define her preferences. *This feedback enriches the individual preference model (IM) and the group preference model (GUM).*
2	Individual	#2	Area #2 is the awareness area. Drag and drop the current recommendation displayed in number 1 to the stack area (number 5). *This action stores a product that is of interest to the user and the remaining users will see it displayed, to keep awareness of the preferences and of how much they like that product by looking at the stars.*
3	Social	#2	Area #2 is the awareness area. Drag and drop the current recommendation displayed in number 1 to the suggestions box of another user placed at the awareness area (number 2). With this action a user draws the attention of other group members.
4	Social	#2	In area #2 user also knows what products the other users are currently viewing and how much she likes it, because the color box contains a 0–5 heart score that denotes how compatible is the product to the user. *A user is able to select the product, which is then moved to area #1.*
5	Social	#4	Area #4 is the suggestion box. These suggestions may come from any of the group members and as a result of a proactive suggestion of the recommendation algorithm. The color boxes contains a 0-5 star score that denotes how compatible is the product to the user. *A user is able to select the product, which is then moved to area #1.*
6	Social	#5	Area #5 is the stack area. It serves as the repository of particular products the user is interested in and it is also useful to draw the attention of the other group members over a particular product. The color boxes display a summary of the product and a 0-5 heart visual score to denote how compatible is the product to the user. *A user is able to select any product, which is then moved to area #1 to view it in detail.*
7	Individual	#3	Area #3 is the waste basket area. It is used for the disposal of products from the suggestion box area (#4) or the stack area (#5). *This action removes a product from the stack and from the suggestion box area.*

algorithm. The recommendation algorithm suggests a product to the whole group when one or more products exceeds a certain critical *compatibility threshold* with respect to the group preference model. The compatibility score was defined originally in Reference [45] as the percentage of critiques in the user model that a product satisfies. As the previous area, each product is identified by the border color and shows few features including a compatibility score with the current user. Furthermore, the option of clicking on the product is also available, to take a look at it in the individual interaction area.

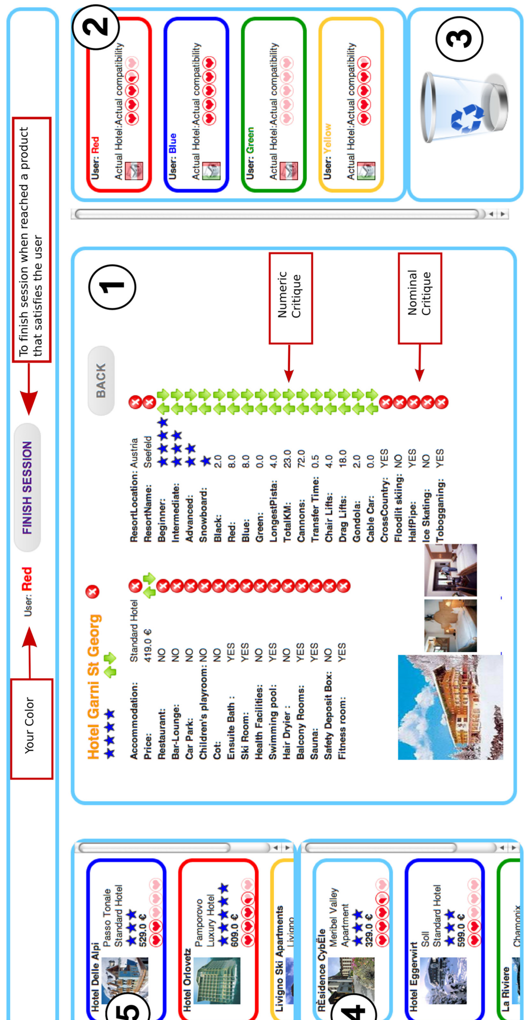


Fig. 2. The main interface of the framework with a skiing package view.

Another form of social interaction is the *stack area*, shown in Figure 2, number 5. It serves as repository of particular holiday recommendations the user is interested in and it is also useful to draw the attention of the other group members over a particular product. The stack stores summaries of the user's recommendations, as well as displaying information related to the compatibility with the group. Each product recommendation appears boxed with the color of the user that added it to the list and shows a summary of its features (in the skiing domain, *hotel name*, *resort name*, *kind of resort*, *number of stars*, and *price*). In addition to this content, a measure of compatibility between the current user and the item appears through a visual score of 0–5 hearts (with 5 indicating a perfect item). When users detect an attractive product in this area, they can open it in the individual interaction area by performing a simple click on it.

The interface considers an area for removing products that the user is not interested in anymore; this area is called the *waste basket* area (see Figure 2, number 3). It is used for the disposal of

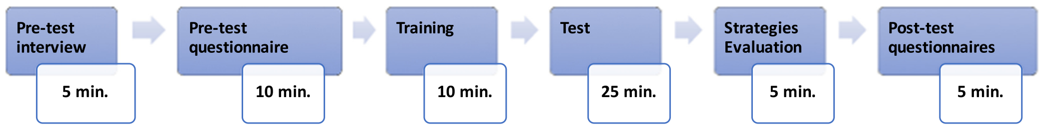


Fig. 3. Phases of the test protocol with their average duration time.

products from the *suggestions box area* or the *stack area*. This functionality is activated when the user performs a drag and drop from one of these areas to the waste basket area.

Finally, the user finishes her session using the *Finish session* button, shown in Figure 2 at the top of the interface. At this point, the user receives a new web page with the information of the products that the group members have added to the stack. The user has to choose one of them as her final decision product. Once all the users have selected their final decision product, an automatic consensus [48] is performed and a product is sent to each member as the final chosen product for the group.

Considering all the interaction modalities, it becomes natural to use the behavior implicitly collected by the gCOACH interface to define roles (i.e., leader or collaborative user) among the group members, so that this can influence the conversational group recommendation algorithm. Note that the interaction modalities let us know the social interactions, the individual actions, and how users respond to other members' actions.

5.4 Session Workflow

We used a summative or quality assurance test, which is usually performed in medium-near to end stages of development. This type of test is adequate for prototypes that already incorporate the major part of the required functionalities and focuses on gathering both qualitative and quantitative data [3]. The test was conducted by a moderator and an observer. The former guided users (if needed), introduced the test, gave users the pre-test interview, the pre-test questionnaire, and the post-test questionnaires. The latter took notes during the test. It is important to highlight that users can ask the moderator questions if necessary during any time of the test.

The test was carried out synchronously in a computer lab at the University Arturo Prat (Chile) for two main reasons: (i) we pursue an effective monitoring of the experiments, since it is quite difficult to obtain volunteers in a group activity, and (ii) gCOACH framework only works in lab environments and it is not fully available for operating environments, because it lacks the module to manage more than a group at the same time. In the test, users were joined in groups based on the relationship ranging between participants, as we have described in Section 5.1. It is important to remark that the group of users were sitting separately in the lab with the aim that they only may interact through the gCOACH interface.

The test protocol consisted of six stages, described below and summarized in Figure 3:

(1) **Pre-test interview:** In this stage, the moderator welcomed the group of users, briefly explained test objectives, annotated personal information such as name, gender, educational level, degree of relationship with other group members, and age. Finally, users were asked about their previous experience with ski vacations and recommender systems.
(2) **Pre-test questionnaire:** In this stage, users were asked to fill out a Web form that contains a questionnaire consisting of 30 questions. The pre-test questionnaire is based on the **Thomas-Killmann Conflict Mode Instrument (TKI)** [29]. We use TKI because it is one of the most popular bargaining style assessment tools. TKI has been used for a variety of purposes, mostly to aid in teaching activities. Here, we concentrate on analyzing the conflict behavior, to evaluate whether the bargaining styles are related to the roles of the group

members. Briefly, TKI [29] defines that in conflict situations an individual's behaviour can be described along two dimensions: (1) assertiveness and (2) cooperativeness. These two dimensions can then be used to define five different modes for responding to conflict situations, such as reach a consensus agreement in a group recommender system: (1) **competing** is assertive and uncooperative; (2) **accommodating** is unassertive and cooperative; (3) **avoiding** is unassertive and uncooperative; (4) **collaborating** is both assertive and cooperative; and (5) **compromising** is moderate in both assertiveness and cooperativeness.

(3) **Training:** During this stage, users were freely navigating on the web interface. Users were asked to locate a predefined product using individual or social interaction modalities. The training stage finished when users discovered this product. It is important to remark that we paid special attention to the design of this stage, to make users aware of the different interaction modalities that the framework enables. The moderator gave a summary sheet of the different interaction modalities of the group interface. In addition, the moderator was in charge of guiding users and users can ask the moderator questions if necessary.

(4) **Test:** In this phase, users performed, without guidance, a test task that consisted of selecting a product that best satisfies the group preferences for going skiing together. To this end, users were asked to navigate, communicate, and provide suggestions with the aim of finding a consensus in the group to purchase a product. However, users were free to finish the search process once they had found a product that best satisfied their preferences. Among the products in the stack, the user selects the preferred one. When all users finished the recommendation process, the system showed for each consensus strategy a recommended product. During the task, a computer recorded the test session and the observer made annotations.

(5) **Strategies evaluation:** In this stage, users were asked to rate the recommended product for each consensus strategy by means a Web form. Concretely, users answered using a star rating. In particular, the highest number of stars indicates the best quality of the recommendation for each consensus strategy. Users were unaware of the strategy that they were evaluating, they just received the item chosen as the final product by each consensus-based strategy and they rated it. The strategies used in this evaluation are *mean*, *multiplicative*, *completeness* that only use the individual interaction to generate a recommendation, and *collaborative mean*, *collaborative multiplicative, collaborative completeness*, *maximize leader satisfaction*, and *leader selection*, which are the proposed ones that use both the individual and collaborative interactions.

(6) **Post-test questionnaires:** Finally, users were asked to fill out a Web form that contains two questionnaires: a collaboration and a satisfaction questionnaire. In the collaboration questionnaire, see Table 3, users answered Q1 and Q2 using a 4-point Likert scale, where 1 corresponds to user "red," 2 to user "blue," 3 to user "green," and 4 to user "yellow," whereas in Q3 and Q4 users answered the questionnaire using a 5-point Likert scale, where 1 correspond to "strongly disagree" and 5 to "strongly agree." The post-test collaboration questionnaire is founded on the work of Krogel et al. [31, 32].

Collaboration among users depends on two factors, first, users are willing to collaborate, and second, the interface has to facilitate such collaboration. For that reason, we were also interested on knowing how the interface helped the users and their satisfaction with it. Following this idea, users were asked to fill out a Web form that contains a satisfaction questionnaire (see Table 4) consisting of six questions. Users answered the questionnaire using a 5-point Likert scale, where 1 correspond to "strongly disagree" and 5 to "strongly agree," with the exception of question 6 that refers to each special area in the interface. It is important to remark that our satisfaction questionnaire follows the methodology proposed in References [30, 43, 46], which is widely used in recommender systems evaluations.

Table 3. Post-test Collaboration Questionnaire

Question Number	Statement
Q1 (Collaborative)	Who do you think has been the most collaborative user?
Q2 (Leader)	Who do you think has been the leader within the group?
Q3 (Friendship)	My friendship with another user in the group affected my final choice?
Q4 (Leader Relationship)	I think that my final choice of leader is consistent with the interactions the leader has made.

Table 4. Post-test Satisfaction Questionnaire

Question Number	Statement
Q1 (Learnability)	The interface provides an adequate way for me to express my preferences.
Q2 (Learnability)	The recommender's interface provides sufficient information.
Q3 (Learnability)	I became familiar with the recommender system very quickly.
Q4 (Usefulness)	It is easy for me to inform the system if I dislike/like the recommended item.
Q5 (Satisfaction)	The interface let me know easily where are the rest of my teammates at all times.
Q6 (Attention)	I paid more attention to the recommendations from: 1. Stack, 2. Suggestions, 3. Awareness, 4. No Special place.

5.5 Evaluation Metrics and Tests

In our experiments, we begin by analyzing the leadership and collaboration among members of the group. In particular, first, we analyze the users' perception of leadership and collaboration within the group. Each user makes a vote for her perception of the leader and another vote for her perception of the most collaborative user. The leader and the most collaborative user/s will be the group member that receives the maximum number of votes. Second, we evaluate the accuracy of the predicted leader score per user at each group. In particular, we use an accuracy metric defined in Equation(19):

$$acc = \frac{correctedPredictions}{totalPredictions}, \tag{19}$$

where the *correctedPredictions* is the number of groups where the user that obtains the highest leader score is also the most voted one by the live-users in the post-questionnaire and the *totalPrediction* term is the total of groups analyzed.

However, for evaluating the consensus-based strategies, we computed the mean rating for each one. Recall that users rated with one to five ratings (i.e., decimal points were allowed), the recommendations suggested by each consensus strategy in a Web form. Note that a rating expresses how much the user is satisfied with the final decision.

With the aim of demonstrating that the introduction of collaboration and leadership in a consensus-based strategy significantly outperforms baseline strategies, we applied statistical tests. Specifically, we have used Friedman [21] and Bonferroni-Dunn [17] tests. These tests are specialized procedures for testing the significance of differences between multiple means and they also control the multiple hypothesis testing problem that is usually present in a pair t-test. For this reason, given that we were testing eight consensus strategies, the use of Friedman and its corresponding Bonferroni-Dunn test was more appropriate.

To be able to apply the Friedman test, we first computed the *mean rank* (r) of each strategy over in all groups. In particular, the evaluation considers $k = 8$ strategies (i.e., Individual Mean, Individual Multiplicative, Individual Completeness, Collaborative Mean, Collaborative Multiplicative, Collaborative Completeness, Leader Selection, and Maximize Leader Satisfaction) and $N = 17$ different groups for the tests. We ranked alternative strategies, for each group, following the practice of Reference [21]. The one that attains the best performance (i.e., in our case the best rating) is ranked 1, the second best is ranked 2, and so on. In case of ties, average ranks are assigned. Then, to obtain the mean rank of each strategy, we average its rank across all groups. Second, we applied the Friedman test to analyze whether the differences between algorithms is statistically significant. If the null-hypothesis of the Friedman test is rejected, then we can proceed with a post-hoc test. In our case, we finally computed the Bonferroni-Dunn test to find out which algorithms were significantly different.

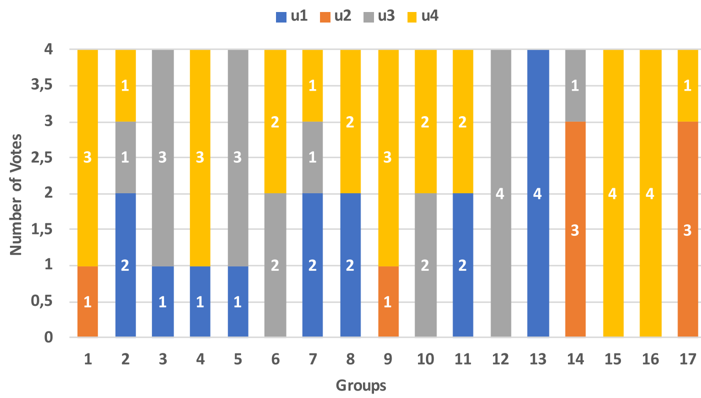
6 RESULTS AND DISCUSSION

In this section, we analyze if our initial hypotheses are confirmed. Recall that the hypotheses are defined in Section 3.4. Accordingly to the hypotheses, we shaped three research questions that have been defined in Section 3.3.

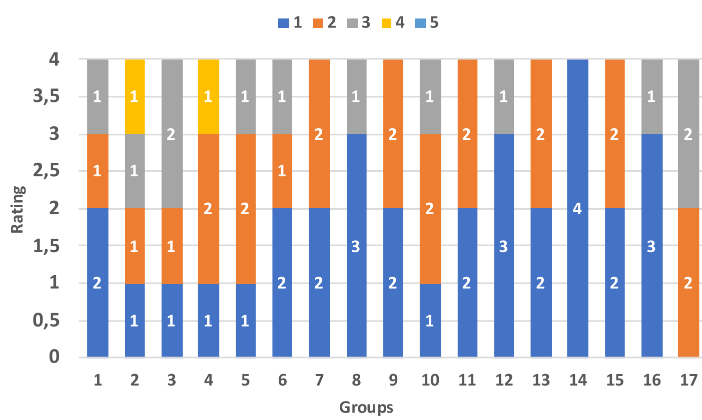
The answers to research questions RQ1, RQ2, and RQ3 will be provided in Sections 6.1, 6.2, and 6.3, respectively. Finally, in Section 6.4, we analyzed the users' perception of the interface, to see which parts of it attracted their attention more, thus generating awareness and allowing them to reach an agreement.

6.1 Analysis of Leadership and Collaboration

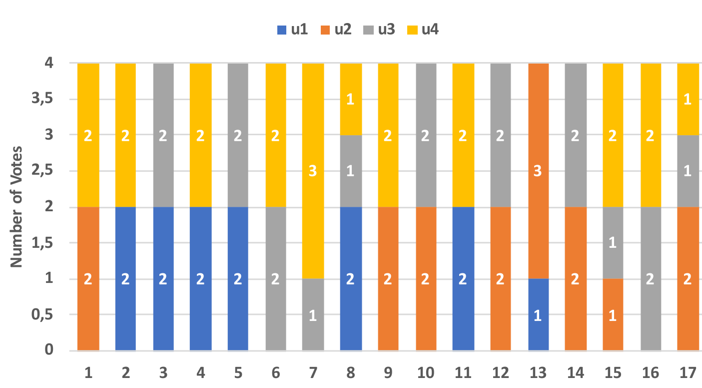
In this section, we begin by analyzing the users' perception of leadership and collaboration within the group. Figure 4 depicts the results for the different 17 groups of 4 users. For each test group, we requested the participants to define the leader and the most collaborative group member (see Q1 and Q2 in Table 3). The bar charts in Figures 4(a) and 4(c) show the number of votes that each user received. The color covering the largest area in a bar corresponds to the user considered as the leader of the group or the most collaborative of the group members. For example, in Figure 4(a), the leader in group 1 is u_4 with 3 votes, whereas in group 2 the leader is u_1 with 2 votes. Another observation is that not all group members are collaborative or leaders. Besides, some groups finish in a draw, as shown in groups 6, 8, 10, and 11 in Figure 4(a). In these groups there are two leaders, because both of them receive two votes. We consider that in groups of four members it is quite easy to finish in a draw. However, note that from a total of 17 groups, users clearly perceive in 13 groups out of 17 that one of the members is the leader. Moreover, in groups 12, 13, 15, and 16 there is a sole leader defined by the users' votes. At first sight, one may think that this is largely due to the relationship among members, as depicted in the third column of Table 1. However, this is not the single rule valid for all the groups. For example, group 8 has the highest relationship among members and they finish in a draw when defining the leader. We also requested participants to define if their final choice has been based on their friendship (Q3 in Table 3). In Q3, the mean rating is 1.85 in a 5-point Likert scale (inverse scale for that question). In fact, 76% of participants think that their friendship with another user in the group did not affect their final choice, so they rated this question with less than 3 points (see Figure 4(b)). Moreover, Figure 4(b) shows that only two participants rated Q3 with 4 points and neither of them rated with 5 points. In addition, note that in Figure 4(b) from a total of 17 groups in 15 of them users rated with less than or equal to 3



(a) Users' votes for the leader user



(b) Users' ratings for the friendship among them at each group



(c) Users' votes for the most collaborative user

Fig. 4. Analysis of Leadership and Collaborative users within the group.

Table 5. Leader Analysis that Includes the Predicted Leader Score Per User at Each Group, the Leader Predicted by the Recommender, and the Leader Defined by the Live-users

	Leader Score					
Group	u_1	u_2	u_3	u_4	**Leader/s Predicted**	**Users' Leader/s**
1	0.0	0.8	0.0	1.0	u_4	u_4
2	1.0	0.2	0.7	0.7	u_1	u_1
3	0.7	0.9	1.0	0.5	u_3	u_3
4	0.9	0.7	0.0	1.0	u_4	u_4
5	0.5	0.4	1.0	0.0	u_3	u_3
6	0.0	0.0	1.0	1.0	u_3, u_4	u_3, u_4
7	1.0	0.0	0.3	0.4	u_1	u_1
8	0.6	0.0	0.3	1.0	u_4	u_1, u_4
9	0.0	0.4	0.0	1.0	u_4	u_4
10	0.7	1.0	0.8	1.0	u_2, u_4	u_3, u_4
11	0.8	0.0	0.0	1.0	u_4	u_1, u_4
12	0.0	0.0	1.0	0.0	u_3	u_3
13	1.0	0.0	0.0	0.0	u_1	u_1
14	0.0	1.0	0.2	0.0	u_2	u_2
15	0.0	0.0	0.0	1.0	u_4	u_4
16	0.2	0.0	0.0	1.0	u_4	u_4
17	0.0	1.0	0.0	0.0	u_2	u_2

points. It is important to highlight that participants with the greatest relationship score in Table 1 (i.e., groups 8, 13, and 17) evaluated with 3 or less point the friendship influence in their final choice. In Q4 (see Table 3), the mean rating is 4.35 and, among participants, 88% of them perceive with more than 3 points that their final choice of leader is consistent with the interactions they have had with the leader. To sum up, our analysis depicts that the leader choice is not based on the relationship or friendship among teammates, as observed in Figure 4(b), and it is mainly based on the users' interactions.

However, Figure 4(c) shows that users have more variability in defining the most collaborative one. From the 17 groups, 12 of them finish in a draw. It is interesting to see that the leader is also one of the most collaborative members of the group. This enforces our approach, since we have considered the leader as the most collaborative user in a group, see Equation (11).

To analyze RQ1, Table 5 details the *leader* score obtained by each member of the group and the chosen leader/s by the recommender and the participants in our experiments. This *leader* score ranges in the interval [0,1], being 1 the score of the leader/s and 0 the minimum score for those users that have not suggested products to the remaining group members. For this reason, there are some users without a leader score. By analyzing the leader score per user at each group, it can be seen that the leader is clearly inferred by the leader score, defined in Equation (11). For example, u_4 is the most voted user in the group 1 (see Figure 4(a) and the last column of Table 5) and this user also reached the highest predicted leader score (see third and fourth columns in Table 5).

It is worth mentioning that the leader of all groups in Table 5 corresponds to the leader defined by the users in our post-questionnaire that can be seen in Figure 4(a) and the last column of Table 5. Thus, the predicted leader score accuracy, defined in Equation(19), is 100% in defining the leader of the group. The user obtaining the largest leader score is the most voted one, in 17 of 17 groups. Note that in those cases where several users finish in a draw, we considered as the leader anyone

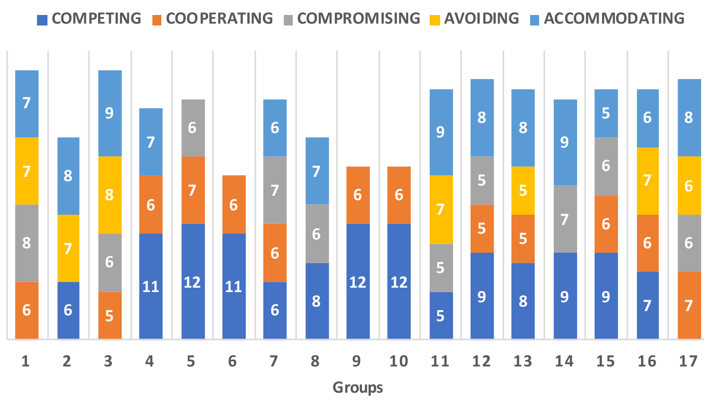


Fig. 5. Analysis of Leadership and Collaborative users within the group.

of them. Nonetheless, as shown in Table 5, the groups that finished in a draw with the users' votes depict a large leader score for all the group members (see for example, groups 6 or 11).

Additionally, the rest of the leader scores depicted in Table 5 correspond to the collaborative users defined by our live users in Figure 4(c). Moreover, when users voted for more than two perceived leaders in a group, in all of these occasions the perceived leaders are predicted with the best scores. For example, the most voted users in the group 2 (u_1, u_3, and u_4), obtained the highest predicted score with values of 1, 0.7, and 0.7, respectively. This means that our predictions also allow us to detect a ranked list of possible leaders in a group. Thus, considering our experiments and the analysis of our results, the answer to our first research question (RQ1) is that we can efficiently infer the leader within a group.

6.2 Analysis of Conflict Styles of Leaders

We have also been interested in analyzing whether the conflict styles are related to the leadership. Figure 5 shows the values of the different conflict resolution styles of the leaders. Recall that users fill in a pre-test questionnaire consisting of 30 questions, as defined in Reference [29]. According to their responses, we extract a score for each mode of resolving conflicts: competing, accommodating, avoiding, collaborating, and compromising. The maximum score for any style is 12 and the total aggregate score is 30. A score higher than 6 on any style indicates a preference of the user for that mode. Since a score lower than 6 indicates relative neglect, as defined in Reference [29], we accordingly remove those styles with a score below 6 in the graph to denote which styles are prominent in a leader.

Figure 5 plots the modes of the leader of each group. In particular, Figure 5 depicts that there is not a predominant mode of conflict style in the leader analysis. However, it can be observed that the competing, cooperating, and accommodating styles are the most common ones for the leaders, covering 67% of the leader styles. Concretely, the leaders seem to influence the rest of the group, because the majority of them have competing (26%) and cooperating (20%) styles. It means that the detected leaders are more assertive in their behavior within the group. Moreover, the score reached by the cooperating and accommodating (21%) styles also show that the leaders have also a high degree of cooperativeness within the group. The second research question (RQ2) aims at analyzing whether we can characterize a leader with a psychological conflict style mode. Considering our experiments and the analysis of our results, the answer to this research question is that there is not a prominent conflict style mode in a group leader.

Table 6. Users' Mean Rating Per Consensus-based Strategy

Strategy	Mean Rating	Std. Deviation
Individual Mean	3.70	1.87
Individual Multiplicative	3.63	1.70
Individual Completeness	3.68	1.63
Collaborative Mean	4.01	1.73
Collaborative Multiplicative	3.64	1.49
Collaborative Completeness	4.03	1.45
Leader Selection	4.26	1.66
Maximize Leader Satisfaction	4.29	1.23

6.3 Analysis of Consensus-based Strategies

However, we have evaluated all the consensus-based strategies with the live users. Recall that users rated the recommended product for each consensus strategy. We would like to analyze if there are differences among well-known individual strategies and those that integrate collaboration. In particular, Table 6 shows the users' mean rating for each strategy. It is important to highlight that all the strategies proposed obtained more than 3.63. Moreover, as we can see in Table 6, the proposed strategies that use both the individual and collaborative models (i.e., Collaborative Mean, Collaborative Multiplicative, Collaborative Completeness, Leader Selection, and Maximize Leader Satisfaction strategies) obtained better results, with an average of 4.05 points, than the ones that only use individual models (i.e., Individual Mean, Individual multiplicative, and Individual Completeness strategies), which reached in average 3.67 points. In addition, the strategies that maximize the interest of the leader (i.e., Leader Selection and Maximize Leader Satisfaction strategies) reached the best results with an average rating of 4.27. Analyzing these best strategies, we can observe that users rated with a better score the strategy that use all products stored in the stack (i.e., the maximize leader satisfaction strategy with 4.29 points) in front of the strategy that only use the stack of the leader (i.e., the leader selection strategy with 4.26 points).

To demonstrate that the hypothesis concerning the introduction of collaboration in a consensus-based strategy significantly outperforming baseline strategies, and to respond to the research question 3, we applied the Friedman [21] and Bonferroni-Dunn [17] tests.

First, we applied the Friedman test to analyze whether the difference between algorithms is statistically significant. In particular, we applied the Friedman test and the p-value obtained is equal to 44.897; since it is greater than 0.05, we can reject the null hypothesis. Once we investigated the non-randomness of our results, we computed the Bonferroni-Dunn test to find out which algorithms were significantly different. In our case, when comparing eight algorithms with a critical value $\alpha = 0.05$, we obtained a value of $q_{0.05} = 2.69$ in a two-tailed Bonferroni-Dun test. We obtained a critical difference value of $CD = 2.26$.

The Bonferroni-Dunn test results are illustrated in Figure 6. In this graph, diamonds represent the mean ranks of each strategy and the vertical lines across diamonds indicate the "critical difference," CD. Basically, the efficiency of the two algorithms is significantly different if their vertical lines are not overlapping. For example, it can be seen that Leader selection and Max Leader Sat. performed significantly better than the baselines (Mean, Multiplicative, and Completeness). Concretely, when we compare the Leader selection with the Mean, Multiplicative, and Completeness strategies, we obtained p-values of 0.000028, 0.000026, and 0.000055, respectively. However, the p-values obtained between the Max Leader Sat. and the Mean, Multiplicative and Completeness strategies were 0.000022, 0.000021, and 0.000044. Additionally, we can see that Collaborative

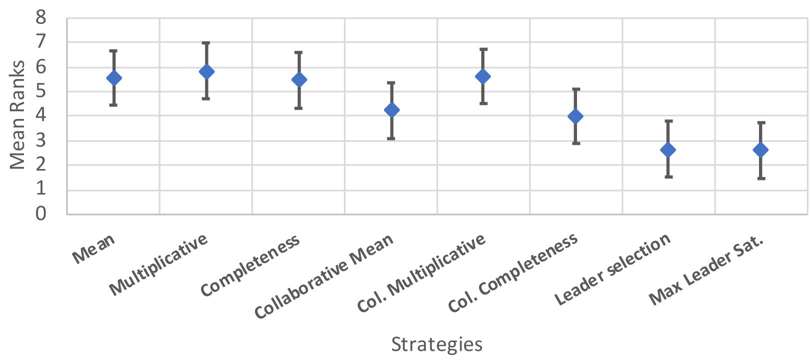


Fig. 6. Application of the Bonferroni-Dunn test to alternative strategies' mean rank of rating.

Mean did not perform significantly better than the baseline Mean ($p = 0.588$) although its mean rank is lower than the Mean strategy. Our proposals that integrate collaborative interactions into consensus-based strategies improved the rating of the baselines (i.e., Mean, Multiplicative, and Completeness), as their mean rank is shorter than them; however, only Leader Selection and Max Leader Sat. are significantly better than the baselines. Among our proposals, Max Leader Sat. obtained the best mean rank, with a value of 2.61, and Leader Selection is very close to it with a mean rank value of 2.64. Note that we can see that the results for both Leader Selection and Max Leader Sat. are significantly better than the individual strategies, with a confidence of 95%. Thus, the response to our third research question (RQ3) is that leadership and collaboration improve the effectiveness of group recommendations.

In summary, from our significance analysis, we conclude that integrating collaboration in the consensus-based strategies improves the rating obtained by the users, which confirms our initial hypotheses. In addition, our results depict that Leader selection and Max Leader Sat. strategies performed significantly better than the baselines, which confirms that the role of the leader is a key point of recommendation generation for groups.

6.4 Analysis of Users' Perception of the Interface

Since collaboration among users depends on their willingness to collaborate and the interface has to facilitate such collaboration, here we analyze the users' perception of the interface, to see to which parts they have paid more attention to be aware, collaborate, and reach an agreement.

Table 7 depicts the results obtained from the satisfaction post-questionnaire (see Table 4). Note that these results are related to the subjective perception of users but are quantitative data that gives us valuable information about users' perception of usefulness of the framework they used. Overall, the quantitative results obtained from the questionnaire are very satisfactory. It is worth noting that all the question obtained, in average, a rating over 4 points in a 5-point Likert scale (see Mean Rating in Table 7).

Considering the learnability of the interface (i.e., questions Q1–Q3), 100% of participants' responses show that the users found the system easy to learn and evaluated this aspect with 3 or more points. In fact, 3.92% of responses were ranked with 3 points (computed as $\frac{6+2+0}{68*3} * 100 = 3.92$), 53.43% correspond to 4 points, and 42.65% received 5 points. With regard to the usefulness, question Q4 answers depict that a 95,59% of the participants evaluated it with 4 or 5 points and a 4.41% of responses were ranked with 3 points. In addition, users' satisfaction with the interface to keep awareness of the group members, responses to question 5, show that a 98.53% of the participants evaluated positively this aspect with 3 or more points.

Table 7. Results of Post-test Questionnaire in a Five-point Likert Scale

Question	1 point	2 points	3 points	4 points	5 points	Mean Rating
Q1	0	0	6	31	31	4.36
Q2	0	0	2	39	27	4.37
Q3	0	0	0	39	29	4.43
Q4	0	0	3	29	36	4.49
Q5	0	1	2	47	18	4.21

Table 8. Users' Attention at the Different Areas of the Interface

Area on the interface	Percentage
Suggestions	64.71%
Stack	13.24%
No special place	11.77%
Awareness	10.29%

Finally, in question 6, we also asked users about the area on the interface that they paid more attention to: the stack area, the suggestion box area, the awareness area, or if there is not a preferred area. We report the results on the users' perception of most useful area in Table 8: 64.71% of users prefer the suggestions area, which denotes that group interaction is highly influenced by the collaboration between users; 13.24% of participants prefer the stack area as its main source of recommendations; there is a 11.77% of users that have not a clear preferences on any area, because they have been looking equally at all of them; finally, 10.29% of users prefer the awareness area as its main source of information about member's activities and for choosing a product.

With this analysis, we can conclude that the users' perception of the usability of the interface is high. Besides, the area of the interface that users paid more attention is the one devoted to suggestions as their main source of information about member's activities and for collaborating. Note that suggestions received in this area come from teammates. This means that group interaction is highly influenced by collaboration as users prefer to observe which products are suggested by their teammates and then select one.

6.5 Discussion

This section summarizes our results with take-home messages and relates them with those obtained in previous literature.

As we have highlighted in the Introduction, in Reference [39], the leader in a group is the user who provided more feedback to the items proposed to the other group members. In that work, no assessment on how the rest of the group perceived the person designated as a leader was made. In addition, 40% of the group felt the final choice did not reflect their preferences. In our case, instead, the item chosen with our notion of leadership achieves the highest rating w.r.t. to the classic non-collaborative strategies. This confirms the assumption we made in the Introduction, that a more complex notion of leadership, where a user does not only have to be active but also needs to propose items that are perceived as valuable by the rest of the group, makes a difference in the perception of the group recommendation.

The integration of a collaborative model based on the social interactions is a powerful tool to implicitly infer the role of each user within the group. Indeed, the user perception of leadership reported in Figure 4(a) clearly corresponds to the leadership predicted by our approach in Table 5.

Since this is the first approach that allows users to interact in a collaborative environment, an important outcome of this work is that a conversational group recommender system is an effective environment that favors interaction, and that our user interface favors learnability, is perceived as useful, and satisfactory for the users. In addition to this, all the different areas in which users can receive recommendations receive attention from the users.

Another important aspect that comes out of our live-user evaluation is that the social interaction activities help to define the true preferences of the group. Indeed, in Table 6 the collaboration-based consensus strategies resulted in a better rating than the classical ones. Additionally, people may change their preferences (or at least their constraints become soft) if the leader is the one that proposes the product. This can also be seen in Table 6, where "Leader Selection" and "Max Leader Satisfaction" are those that obtained the highest scores.

Our results confirm our hypotheses that by monitoring user interactions, we can build a collaborative model and infer useful notions of collaboration and leadership, which can concretely help generating more effectiveness of group recommendations.

The results we obtained distance our work from most of the previous related literature. While in Section 2, we have introduced several approaches that integrate the personality score in the group model, our results show that personality does not reflect neither collaboration nor leadership (recall that Figure 5 highlights that each leader approaches the group with a different conflict style mode). Specifically, Sánchez et al. [49] detected that half of their user base has a strong personality, which had an impact in the way their preferences where valued; nevertheless, our study shows that in an online collaborative settings, personality does not play a significant role in defining who a leader is, which is shaped through behavioral dynamics during the recommendation session. Finally, while previous observational studies, such as Reference [14], highlighted that multiplicative utilitarian is the strategy that does not consider user interactions and is closer to the consensus reached by a group, our results show that individual completeness is the strategy with the highest rating when collaboration is not integrated (see Figure 6). However, the most important aspect to highlight is that our study related to the existing observational ones, since collaboration allows to significantly improve the quality of the group recommendations that are being produced.

Regarding the limitation of our work, our live-user evaluation (and the pandemic scenario under which the study was conducted) limited the amount of groups we could consider. In the future, we plan to extend our study and consider more groups, to get more insights on the role of interactions, collaboration, and leadership in conversational group recommender systems. In addition to this, our evaluation was limited to one application domain (skiing group recommendation); the insights that might emergence from the collaboration in different domains remain open and the consideration of additional scenarios is left as future work. Nevertheless, we believe our framework, since it captures notions of collaboration and leadership from user behavior, can ensure the generalizability of our results besides the skiing domain. Indeed, items usually have a picture and an item description; so our platform can accommodate data coming from any domain, to support group decision-making processes.

In conclusion, we can highlight that leadership in conversational recommender systems is a role that emerges from the behavior of the users in the browsing session. This collaborative process smooths previous personality traits, with strongly collaborative users that are recognized by the whole group as leaders. Leaders can influence the choice of the whole group, without affecting the satisfaction of the rest of the group.

7 CONCLUSIONS AND FUTURE WORK

In this article, we tackled the issue of monitoring user interactions, to infer notions of collaboration and leadership, with the aim to improve the effectiveness of **group recommender systems**

(GRS). To this end, we have modeled social interactions between users, and have implemented a new collaboration model. Moreover, we developed new collaborative-consensus strategies based on the proposed model.

We evaluated our proposals using an online conversational group recommender framework with real users. After the evaluation, users were requested to designate the most collaborative and the leader of the group. The GRS constructs the collaborative model during the interaction and extract two scores, i.e., the collaborative score and the leader score. When compared the leader score against the one proposed (i.e., the most voted) by real users, our results show an accuracy of 100% inferring the leader user(s).

However, the live-user evaluation also demonstrates that the proposed collaborative-based consensus strategies, which use both an individual and collaborative user models, improve the effectiveness of the GRS. Specifically, strategies that use exclusively the leader's selections, significantly improve the classic ones. Additionally, we analyzed the relationship between the conflict style modes of the leader within the group. We can conclude that the leaders seem to influence the rest of the group and they also have a high degree of cooperativeness (defined in two different style modes: accommodating and/or collaborating) and assertiveness (denoted by the competing mode). However, the most important observation has been that there is not a rule in applying these styles to define a leader.

As a future work, we will address an in-depth analysis of pre-existing social relationship between users. Moreover, it will be interesting to analyze in depth the conflict style modes of the whole group and their inter-relationships. Finally, while we considered a conversational approach, other types of algorithms such as collaborative and content-based ones [19] could also be considered in the future.

REFERENCES

[1] Jesús Omar Álvarez Márquez and Jürgen Ziegler. 2016. Hootle+: A group recommender system supporting preference negotiation. In *Proceedings of the 22nd International Conference on Collaboration and Technology Collaboration and Technology (CRIWG'16)*, Takaya Yuizono, Hiroaki Ogata, Ulrich Hoppe, and Julita Vassileva (Eds.), Vol. 9848. Lecture Notes in Computer Science. Springer International Publishing, Cham, 151–166. https://doi.org/10.1007/978-3-319-44799-5_12

[2] Francesco Barile, Judith Masthoff, and Silvia Rossi. 2017. A detailed analysis of the impact of tie strength and conflicts on social influence. In *Proceedings of the 25th Conference on User Modeling, Adaptation, and Personalization (UMAP'17)*, Mária Bieliková, Eelco Herder, Federica Cena, and Michel C. Desmarais (Eds.). Association for Computing Machinery, New York, NY, 227–230. https://doi.org/10.1145/3099023.3099056

[3] D. Bowman, J. Gabbard, and D. Hix. 2002. A survey of usability evaluation in virtual environments: Classification and comparison of methods. *Presence: Teleoper. Virtual Environ.* 11, 4 (2002), 404–424. https://doi.org/10.1162/105474602760204309

[4] Robin D. Burke, Kristian J. Hammond, and Benjamin C. Young. 1997. The findme approach to assisted browsing. *IEEE Expert: Intell. Syst. Appl.* 12, 4 (1997), 32–40. https://doi.org/10.1109/64.608186

[5] Martin M. Chemers. 2014. *An Integrative Theory of Leadership.* Taylor & Francis.

[6] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: Survey and emerging trends. *User Model. User-Adapt. Interact.* 22, 1-2 (2012), 125–150. https://doi.org/10.1007/s11257-011-9108-6

[7] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. ACM, 815–824. https://doi.org/10.1145/2939672.2939746

[8] Ingrid Alina Christensen and Silvia N. Schiaffino. 2014. Social influence in group recommender systems. *Online Info. Rev.* 38, 4 (2014), 524–542. https://doi.org/10.1108/OIR-08-2013-0187

[9] David Contreras, Maria Salamó, and Jordi Pascual. 2015. A web-based environment to support online and collaborative group recommendation scenarios. *Appl. Artific. Intell.* 29, 5 (May 2015), 480–499. https://doi.org/10.1080/08839514.2015.1026661

[10] David Contreras, Maria Salamó, Inmaculada Rodríguez, and Anna Puig. 2018. Shopping decisions made in a virtual world: Defining a state-based model of collaborative and conversational user-recommender interactions. *IEEE Consum. Electron. Mag.* 7 (2018), 26–35. https://doi.org/10.1109/MCE.2017.2728819

[11] Sriharsha Dara, C. Ravindranath Chowdary, and Chintoo Kumar. 2019. A survey on group recommender systems. *J. Intell. Info. Syst.* 54 (Jan. 2019), 271–295. https://doi.org/10.1007/s10844-018-0542-3

[12] Amra Delic, Judith Masthoff, Julia Neidhardt, and Hannes Werthner. 2018. How to use social relationships in group recommenders: Empirical evidence. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP'18)*, Tanja Mitrovic, Jie Zhang, Li Chen, and David Chin (Eds.). Association for Computing Machinery, New York, NY, 121–129. https://doi.org/10.1145/3209219.3209226

[13] Amra Delic, Julia Neidhardt, Thuy Ngoc Nguyen, and Francesco Ricci. 2016. Research methods for group recommender system. In *Proceedings of the Workshop on Recommenders in Tourism co-located with 10th ACM Conference on Recommender Systems (RecSys'16) (CEUR Workshop Proceedings)*, Daniel R. Fesenmaier, Tsvi Kuflik, and Julia Neidhardt (Eds.), Vol. 1685. CEUR-WS.org, New York, NY, 30–37. Retrieved from http://ceur-ws.org/Vol-1685/paper5.pdf.

[14] Amra Delic, Julia Neidhardt, Thuy Ngoc Nguyen, Francesco Ricci, Laurens Rook, Hannes Werthner, and Markus Zanker. 2016. Observing group decision making processes. In *Proceedings of the 10th ACM Conference on Recommender Systems*, Shilad Sen, Werner Geyer, Jill Freyne, and Pablo Castells (Eds.). ACM, New York, NY, 147–150. https://doi.org/10.1145/2959100.2959168

[15] Amra Delic, Julia Neidhardt, Laurens Rook, Hannes Werthner, and Markus Zanker. 2017. Researching individual satisfaction with group decisions in tourism: Experimental evidence. In *Proceedings of the International Conference in Information and Communication Technologies in Tourism (ENTER'17)*, Roland Schegg and Brigitte Stangl (Eds.). Springer, Cham, Berlin, 73–85. https://doi.org/10.1007/978-3-319-51168-9_6

[16] Amra Delić, Thuy Ngoc Nguyen, and Marko Tkalčič. 2020. Group decision-making and designing group recommender systems. In *Handbook of e-Tourism*, Zheng Xiang, Matthias Fuchs, Ulrike Gretzel, and Wolfram Höpken (Eds.). Springer International Publishing, Cham, 1–23. https://doi.org/10.1007/978-3-030-05324-6_57-1

[17] Olive Jean Dunn. 1961. Multiple comparisons among means. *J. Amer. Statist. Assoc.* 56, 293 (1961), 52–64.

[18] Alexander Felfernig, Ludovico Boratto, Martin Stettinger, and Marko Tkalcic. 2018. *Group Recommender Systems: An Introduction*. Springer US. https://doi.org/10.1007/978-3-319-75067-5

[19] Alexander Felfernig, Atas Müslü, Denic Helis, Thi Ngoc Trang Tran, Martin Stettinger, and Ralph Samer. 2018. *Algorithms for Group Recommendation*. Springer International Publishing, Cham, 27–58. https://doi.org/10.1007/978-3-319-75067-5_2

[20] Donelson R. Forsyth. 2012. Social influence and group behavior. In *Handbook of Psychology*, 2nd ed. Vol. 5. American Cancer Society, New York, Chapter 14, 305–328. https://doi.org/10.1002/9781118133880.hop205014 Retrieved from arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118133880.hop205014.

[21] M. Friedman. 1940. A comparison of alternative tests of significance for the problem of m rankings. *Ann. Math. Stat.* 11, 1 (Mar. 1940), 86–92.

[22] Mike Gartrell, Xinyu Xing, Qin Lv, Aaron Beach, Richard Han, Shivakant Mishra, and Karim Seada. 2010. Enhancing group recommendation by incorporating social relationship interactions. In *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work (GROUP'10)*, Wayne G. Lutters, Diane H. Sonnenwald, Tom Gross, and Madhu Reddy (Eds.). ACM, New York, 97–106. https://doi.org/10.1145/1880071.1880087

[23] Francesca Guzzi, Francesco Ricci, and Robin D. Burke. 2011. Interactive multi-party critiquing for group recommendation. In *Proceedings of the ACM Conference on Recommender Systems (RecSys'11)*, Bamshad Mobasher, Robin D. Burke, Dietmar Jannach, and Gediminas Adomavicius (Eds.). ACM, New York, 265–268. https://doi.org/10.1145/2043932.2043980

[24] Marc Guell, Maria Salamo, David Contreras, and Ludovico Boratto. 2020. Integrating a cognitive assistant within a critique-based recommender system. *Cogn. Syst. Res.* 64 (2020), 1–14. https://doi.org/10.1016/j.cogsys.2020.07.003

[25] Michael Hogg. 2010. Influence and leadership. In *Handbook of Social Psychology*. John Wiley & Sons, 1166–1207. https://doi.org/10.1002/9780470561119.socpsy002031

[26] Anthony Jameson and Barry Smyth. 2007. Recommendation to groups. In *The Adaptive Web, Methods and Strategies of Web Personalization*. Lecture Notes in Computer Science, Vol. 4321. Springer-Verlag, Berlin, 596–627. https://doi.org/10.1007/978-3-540-72079-9_20

[27] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2020. A survey on conversational recommender systems. Retrieved from https://arxiv.org/abs/2004.00646.

[28] Epaminondas Kapetanios. 2008. Quo vadis computer science: From turing to personal computer, personal content and collective intelligence. *J. Data Knowl. Eng.* 67, 2 (2008), 286–292. https://doi.org/10.1016/j.datak.2008.05.003

[29] R. Kilmann and K. W. Thomas. 1976. Interpersonal conflict-handling behavior as reflections of Jungian personality dimensions. *Psychol. Reports* 37 (01 1976), 971–980. https://doi.org/10.2466/pr0.1975.37.3.971

[30] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Model. User-Adapt. Interact.* 22, 4-5 (2012), 441–504. https://doi.org/10.1007/s11257-011-9118-4

[31] Julieann Krogel. 2008. *The Group Questionnaire: A new measure of the group relationship*. Ph.D. Dissertation. Brigham Young University–Provo. Retrieved from https://scholarsarchive.byu.edu/etd/1732/.

[32] Julieann Krogel, Gary Burlingame, Chris Chapman, Tyler Renshaw, Robert Gleave, Mark Beecher, and Rebecca MacNair-Semands. 2013. The group questionnaire: A clinical and empirically derived measure of group relationship. *Psychother. Res.* 23, 3 (2013), 344–354. https://doi.org/10.1080/10503307.2012.729868

[33] J. M. Levine and R. S. Tindale. 2015. Social influence in groups. *APA Handbooks in Psychology. APA Handbook of Personality and Social Psychology*, vol. 2 (2015), 3–34. https://doi.org/10.1037/14342-001

[34] Kevin McCarthy, Lorraine McGinty, Barry Smyth, and Maria Salamó. 2006. The needs of the many: A case-based group recommender system. In *Proceedings of the 8th European Conference on Advances in Case-based Reasoning*. Springer Verlag, Berlin, 196–210. https://doi.org/10.1007/11805816_16

[35] Kevin McCarthy, Maria Salamó, Lorcan Coyle, Lorraine McGinty, Barry Smyth, and Paddy Nixon. 2006. Group recommender systems: A critiquing-based approach. In *Proceedings of the 11th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, 267–269. https://doi.org/10.1145/1111449.1111506

[36] Robert R. McCrae and Oliver P. John. 1992. An introduction to the five-factor model and its applications. *J. Personal.* 60, 2 (1992), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6494.1992.tb00970.x.

[37] Lorraine McGinty and James Reilly. 2011. On the evolution of critiquing recommenders. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (Eds.). Springer, Berlin, 419–453. https://doi.org/10.1007/978-0-387-85820-3_13

[38] Felipe Moraes, Kilian Grashoff, and Claudia Hauff. 2019. On the impact of group size on collaborative search effectiveness. *Info. Retriev. J.* 22, 5 (2019), 476–498. https://doi.org/10.1007/s10791-018-09350-9

[39] Thuy Ngoc Nguyen and Francesco Ricci. 2018. A chat-based group recommender system for tourism. *J. IT Tour.* 18, 1-4 (2018), 5–28. https://doi.org/10.1007/s40558-017-0099-y

[40] Thuy Ngoc Nguyen and Francesco Ricci. 2018. Situation-dependent combination of long-term and session-based preferences in group recommendations: An experimental analysis. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC'18)*, Hisham M. Haddad, Roger L. Wainwright, and Richard Chbeir (Eds.). ACM, New York, 1366–1373. https://doi.org/10.1145/3167132.3167279

[41] Mark O'Connor, Dan Cosley, Joseph A. Konstan, and John Riedl. 2001. PolyLens: A recommender system for groups of users. In *Proceedings of the 7th European Conference on Computer Supported Cooperative Work*, Wolfgang Prinz, Matthias Jarke, Yvonne Rogers, Kjeld Schmidt, and Volker Wulf (Eds.). Kluwer, Bonn, Germany, 199–218. https://doi.org/10.1007/0-306-48019-0_11

[42] Jordi Pascual, David Contreras, and Maria Salamó. 2014. Analysis of a collaborative advisory channel for group recommendation, In *Proceedings of the 17th International Conference of the Catalan Association for Artificial Intelligence*, Lledó Museros Cabedo, Oriol Pujol, and Núria Agell (Eds.). *Frontiers in Artificial Intelligence and Applications* 269, Artificial Intelligence Research and Development, 116–125. https://doi.org/10.3233/978-1-61499-452-7-116

[43] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys'11)*. Association for Computing Machinery, New York, NY, 157–164. https://doi.org/10.1145/2043932.2043962

[44] Guilherme Ramos and Carlos Caleiro. 2020. A novel similarity measure for group recommender systems with optimal time complexity. In *Proceedings of the 1st International Workshop on Bias and Social Aspects in Search and Recommendation (BIAS'20) (Communications in Computer and Information Science)*, Ludovico Boratto, Stefano Faralli, Mirko Marras, and Giovanni Stilo (Eds.), Vol. 1245. Springer, 95–109. https://doi.org/10.1007/978-3-030-52485-2_10

[45] James Reilly, Kevin McCarthy, Lorraine McGinty, and Barry Smyth. 2005. Incremental critiquing. *Knowl.-Based Syst.* 18, 4-5 (2005), 143–151. https://doi.org/10.1016/j.knosys.2004.10.005

[46] Francesco Ricci and Quang Nhat Nguyen. 2007. Acquiring and revising preferences in a critique-based mobile recommender system. *IEEE Intell. Syst.* 22, 3 (2007), 22–29. https://doi.org/10.1109/MIS.2007.43

[47] Maria Salamó and Sergio Escalera. 2012. Increasing retrieval quality in conversational recommenders. *IEEE Trans. Knowl. Data Eng.* 24, 10 (2012), 1876–1888. https://doi.org/10.1109/TKDE.2011.116

[48] Maria Salamó, Kevin McCarthy, and Barry Smyth. 2012. Generating recommendations for consensus negotiation in group personalization services. *Person. Ubiq. Comput.* 16, 5 (2012), 597–610. https://doi.org/10.1007/s00779-011-0413-1

[49] Lara Quijano Sánchez, Belén Díaz-Agudo, and Juan A. Recio-García. 2014. Development of a group recommender application in a social network. *Knowl.-Based Syst.* 71 (2014), 72–85. https://doi.org/10.1016/j.knosys.2014.05.013

[50] Lara Quijano Sánchez, Juan A. Recio-García, and Belén Díaz-Agudo. 2011. Using personality to create alliances in group recommender systems. In *Proceedings of the 19th International Conference on Case-based Reasoning (Lecture Notes in Computer Science)*, Ashwin Ram and Nirmalie Wiratunga (Eds.), Vol. 6880. Springer, London, UK, 226–240. https://doi.org/10.1007/978-3-642-23291-6_18

[51] Lara Quijano Sánchez, Juan A. Recio-García, Belén Díaz-Agudo, and Guillermo Jiménez-Díaz. 2013. Social factors in group recommender systems. *ACM Trans. Intell. Syst. Technol.* 4, 1 (2013), 8:1–8:30. https://doi.org/10.1145/2414425.2414433

[52] Chirag Shah and Roberto González-Ibáñez. 2011. Evaluating the synergic effect of collaboration in information seeking. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'11)*. Association for Computing Machinery, New York, NY, 913–922. https://doi.org/10.1145/2009916.2010038

[53] Laure Soulier, Chirag Shah, and Lynda Tamine. 2014. User-driven system-mediated collaborative information retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'14)*. Association for Computing Machinery, New York, NY, 485–494. https://doi.org/10.1145/2600428.2609598

[54] Laure Soulier, Lynda Tamine, and Chirag Shah. 2016. MineRank: Leveraging users' latent roles for unsupervised collaborative information retrieval. *Info. Process. Manage.* 52, 6 (2016), 1122–1141. https://doi.org/10.1016/j.ipm.2016.05.002

[55] Martin Stettinger, Alexander Felfernig, Gerhard Leitner, Stefan Reiterer, and Michael Jeran. 2015. Counteracting serial position effects in the CHOICLA group decision support environment. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI'15)*. ACM, New York, NY, 148–157. https://doi.org/10.1145/2678025.2701391

[56] Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'18)*. ACM, 235–244. https://doi.org/10.1145/3209978.3210002

[57] Marko Tkalčič, Amra Delic, and Alexander Felfernig. 2018. *Personality, Emotions, and Group Dynamics*. Springer International Publishing, Cham, 157–167. https://doi.org/10.1007/978-3-319-75067-5_9

[58] Lucas Vinh Tran, Tuan-Anh Nguyen Pham, Yi Tay, Yiding Liu, Gao Cong, and Xiaoli Li. 2019. Interact and decide: Medley of sub-attention networks for effective group recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. ACM, New York, NY, 255–264. https://doi.org/10.1145/3331184.3331251

[59] Luyan Xu, Xuan Zhou, and Ujwal Gadiraju. 2020. How does team composition affect knowledge gain of users in collaborative web search? In *Proceedings of the 31st ACM Conference on Hypertext and Social Media, Virtual Event (HT'20)*, Ujwal Gadiraju (Ed.). ACM, 91–100. https://doi.org/10.1145/3372923.3404784

[60] Xiaoying Zhang, Hong Xie, Hang Li, and John C. S. Lui. 2019. Toward building conversational recommender systems: A contextual bandit approach. Retrieved from https://abs/1906.01219.

Received May 2020; revised March 2021; accepted April 2021