Feedback on group recommendations

Lara Quijano-Sánchez[1] and Derek Bridge[2]

[1] Department of Software Engineering and Artificial Intelligence, Universidad Complutense de Madrid, Spain
[2] Department of Computer Science, University College Cork, Ireland
lara.quijano@fdi.ucm.es, d.bridge@cs.ucc.ie

Abstract. This is a discussion paper on the subject of group recommender systems. In the recent past, we have built such a recommender system, *HappyMovie*, and we have used variants of it in a number of experiments. In the light of our experience, we look at the the kind of feedback users might give to a group recommender, informed also by new results from a survey that we conducted. We conclude with ideas for the development of the next generation of group recommender systems.

1 Introduction

Recommender Systems use inferred preferences to suggest to their users items that the users might like to consume. Group Recommender Systems do the same, but they recommend items to a group of users, where the group intends to consume the items together.

Case-Based Reasoning (CBR) has a long history of contributing to recommender systems [2]. Most simply, we can build a case-based recommender system where the cases represent the items (e.g. products) and the CBR application recommends cases that are similar to the user's partially-described preferences. More interestingly, the cases in the case base can instead describe the experience of consuming recommended products [12].

We have built a group recommender system for movies. We have also built a variant of our group recommender that uses CBR in the way described at the end of the previous paragraph. We briefly describe our group recommender and this case-based variant in Section 2.

In the course of developing these recommender systems, we have uncovered a number of perspectives on the kind of feedback that group recommender systems might seek, which we present in Section 3. To make this more concrete, we ran a group recommender system experiment with real users and administered a questionnaire to the participants. We describe the experiment and the results of the questionnaire in Sections 4 and 5. We conclude in Section 6 with ideas for the development of the next generation of group recommender systems.

2 Group Recommender Systems

Commonly, group recommender systems aggregate predicted ratings for group members [4]. First, a single-person recommender system predicts each group member's rating for each candidate item. This might be done, as it is in our *HappyMovie* group recommender system, using a standard user-based, nearest-neighbours collaborative filtering approach. Next, the recommender aggregates the ratings, e.g. by taking their maximum or their average. Finally, it recommends the candidate items that have the highest aggregated predicted ratings.

There are many possible variations on this common approach. Our *HappyMovie* system, for example, applies a function to each predicted rating *before* aggregation [11]:

- On registration with *HappyMovie*, users take a personality test whose results are converted into a personality score between 0 and 1, where 0 means a cooperative person and 1 means a selfish person [15]. A user's predicted rating will count for more in the aggregation if her personality score is higher than that of the other group members.
- After registration, the strength of connection ('trust') between pairs of users is mined from social network data. A person's predicted ratings are pulled towards the opinions of the other group members to a degree based on their strength of connection [3].

In [13], we presented a variant of *HappyMovie* that uses CBR: its aggregation of predicted ratings is a lazy and local generalization of the behaviours captured by the neighbouring cases in the case base. First, it uses a user-based, nearest-neighbours collaborative filtering approach to predict each group member's rating for each candidate item. Next, it retrieves cases, i.e. past group recommendation events, that involve groups that are similar to the active group. Case retrieval uses a user-user similarity measure, and, as a by-product, it aligns each member of the active group with a member of the group in the case. The similarity measure compares group members on their age, gender, personality and ratings and the degrees of trust between members of each group. Then, it reuses each case that is retrieved: the contributions that each group member made in choosing the selected item are transferred to the corresponding member of the active group. This is done by scoring the new candidate items by their item-item similarity to the selected item. In this way, the retrieved cases act as implicit models of group decision-making, which are transferred to the decision-making in the active group. Finally, it recommends the candidate items that have obtained the highest scores.

3 Feedback to Group Recommender Systems

Suppose we have a group recommender; for concreteness, suppose it recommends movies. Consider the scenario where the recommender recommends a movie to a group, the group accept the recommendation, they see the movie together, and some or all of the group members come back and provide explicit feedback in the form of ratings. What sort of feedback should the recommender solicit?

3.1 Actual ratings

Like conventional recommender systems, most group recommender systems ask each user how much she likes the movie, e.g. as a star-rating on a five point scale. User-movie ratings are the most important (and often the only) form of *training data* for collaborative recommender systems. The additional training data may improve single-user predictions. And, since most group recommender systems work by aggregating single-user predictions, this in turn may improve group recommendations. The assumption is that the better the predictions, the better the recommendations.

3.2 User satisfaction with the recommendation

But, even if prediction accuracy is high, it does not follow that recommendation quality will be high. That also depends on how successful the aggregation is. For example, if a user watches a recommended movie in a group and later gives it a low rating, this does not mean that the group recommender has done a poor job. It may even be that the group recommender predicted that this user would give a low rating. But the movie was recommended nonetheless, as it was judged to be the one that best reconciled the different tastes of the group members: sometimes people have to lose out if the recommender is to reach a decision at all; sometimes people lose out to group members who have special priority such as children or members with disabilities; sometimes the preferences of a user who was favoured on a previous occasion may, in the interests of fairness, be weighted lower on a subsequent occasion [14].

So there is a separate dimension that can be measured: user satisfaction with the recommendation. For example, a user who dislikes the movie (gives it a low rating) may nevertheless be satisfied with the recommendation, especially if she appreciates that it has been necessary to balance conflicting interests. Her satisfaction might be all the greater if she has a more accommodating (less selfish) personality type, or if the recommendation better matches the tastes of group members with whom she has stronger connections (so-called contagion and conformity effects [9]). A father who takes his children to the cinema provides one such example: if his children like the recommendation, his own satisfaction with the recommendation may increase.

Additionally, *expectations* can influence satisfaction [9], even in single-user recommenders, and these can be influenced to some extent through explanations (e.g. "None of this week's movies is a good match to your preferences. The one I'm recommending is the best of a poor crop."). This may be even more important in group recommenders where the trade-offs that have been made can be explained.

3.3 The group experience

But there is yet another dimension to group movie-going which goes beyond both whether each member liked the movie (their rating) and their satisfaction with the recommendation. There is what we might call the *experience as a whole* (or just *the experience* for short).[3] Although the movie might be one that a group member would not choose for herself, she may still have had an enjoyable time. She may not have liked the movie; she may not have been satisfied with the recommendation (e.g. in the way that it traded-off her preferences against those of other members of the group), but watching it with her friends was still fun. Indeed, it might even be the case that the majority of the group thought a movie was terrible but they may still have enjoyed watching the movie with these friends, e.g. perhaps its awfulness provoked hilarity or heated discussion. The father watching a movie with his children may have had a great time, and this is distinct from, although not wholly uncorrelated with, his movie rating and his satisfaction with the way the recommendation traded-off group interests. The same is true of most consumption done in groups, e.g. dining out together, making excursions together, and so on —the quality of the experience is not necessarily related to what each user thought of the item, nor the user's satisfaction with the recommendation.

It is also possible that different members of the group may evaluate the group experience in different ways. For example, the heated debate that ensued from a controversial movie may be perceived by one group member to have been exhilarating but perceived by another to have been uncomfortable. On the whole, however, we probably expect some agreement about the group experience due to the contagion and conformity effects mentioned earlier [9].

4 *HappyMovie* Experiment

In an effort to explore these issues further, we ran an experiment with real users. Sixty students from a masters-level Artificial Intelligence course participated. They were between 20 and 26 years' old. Twenty-three were female (38.3%); thirty-seven were male (61.6%). Individually, each student completed a Personality Survey, which used TKI's Alternative Movie Metaphor [15]: for each of five different dimensions of personality, we showed the student two well-known movie characters whose personalities oppose each other along that dimension; the student selected the member of the pair with which she most identified. The result is a numeric score in $[0, 1]$. In essence, a value of zero is a very cooperative person and a value of one is a very selfish person. Each student also completed a Preferences Survey: we asked them to rate 70 well-known movies using a five-point rating scale. *HappyMovie* uses these ratings for its collaborative filtering. Finally, the strength of connection ('trust') between pairs of users was mined from Facebook interactions.

[3] We are not referring here to the user experience that comes from engaging with the software [5]; we are referring to the experience of consuming (in our case, in a group) the recommended items.

We formed 20 groups, each comprising three students.[4] Each group used *HappyMovie* to create a group event —an outing to the cinema together; they received three movie recommendations from *HappyMovie* —the three that the recommender decided were best for the group, from a listing of current movies; and they agreed on one of the recommended movies —the one that their group would go to see. We asked them to imagine going to the cinema to watch that movie with the members of their group.

Then, individually and independently they answered a questionnaire of eight questions.[5] The first seven questions were about the movie that they had selected:

1. Give your personal rating for this movie (0 for a movie you really disliked, up to 5 for a movie you really liked).
2. Give the rating that you think your friend 1 in the group will give to this movie (0 if you think s/he really disliked it, up to 5 if you think s/he really liked it).
3. Give the rating that you think your friend 2 in the group will give to this movie.
4. Evaluate the enjoyability of your experience of watching this movie with your group (0 for a really bad experience, up to 5 for a good experience — where you had a great time together).
5. Evaluate the enjoyability of the experience that you think your friend 1 in the group will have by watching this movie with your group.
6. Evaluate the enjoyability of the experience that you think your friend 2 in the group will have by watching this movie with your group.
7. Out of the listing of current movies, do you think that this would have been your choice if you had to go to the movies together in reality — without using *HappyMovie* (0 for 'No, we would have never chosen this movie', up to 5 'Yes, we would have definitely chosen this movie').

The eighth question asked a more general question about recommendations:

8. When you go to the movies with a group of friends, what do you value most about a recommendation? Order the options by importance (most important first):
   (a) That the movie was a good movie —in terms of quality.
   (b) That you personally enjoyed the movie.
   (c) That you and your friends had a good experience watching the movie.
   (d) That the recommended movie was the one that you would have chosen as a group.

These relate to the discussion in the previous section in the following way: option (b) is related to movie rating (Section 3.1); option (c) is what we called the group experience (Section 3.3); and option (d) is about user satisfaction with the recommendation (Section 3.2). Option (a) is an 'objective' notion of quality.

[4] Three was the average group size reported by 105 movie-goers in a poll that we conducted [10].

[5] We ran the experiment with students whose first language was Spanish. The questions that we show here are paraphrases into English of the Spanish questionnaire.

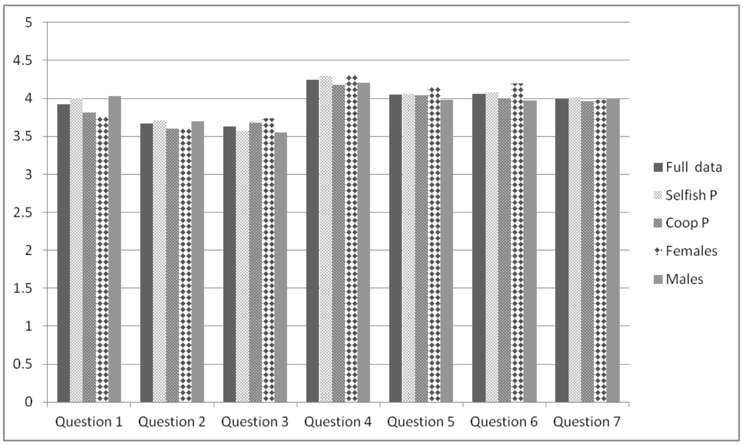

Fig. 1. Average rating by user group of responses to questions 1–7

5 *HappyMovie* Experiment Results

For analysis of the results of the questionnaires, we consider five types of user:

Full data: all sixty users;
Selfish P: the thirty-five users with a more selfish personality, i.e. users whose TKI personality score is no less than 0.6;
Coop P: the twenty-five users with a more cooperative personality, i.e. users whose TKI personality score is less than 0.6;
Females: the twenty-three females; and
Males: the thirty-seven males.

A background observation is that the male students tended to have higher TKI personality values (average 0.68784), implying more selfish personalities, whereas the female students had a lower average TKI personality value (0.46052), implying less selfish personalities.

The results for the first seven questions are in Figure 1. We can conclude:

– On average, these users rate the group experience more highly than they rate the movie (compare Questions 4 and 1), and they think their friends will do the same (Questions 5 & 6 versus 2 & 3).
– On average, these users give higher ratings to the selected movie (Question 1) than they think their friends will give to the movie (Questions 2 and 3). Similarly, their rating of the experience of seeing the movie with these friends (Question 4) is higher than what they think their friends' ratings of the experience will be (Questions 5 and 6). So they feel that the recommender has favoured them, or that they have 'won' in the decision about which movie the group will go to see. This raises the question of whether users tend to rationalise decisions even when the decision goes against them.

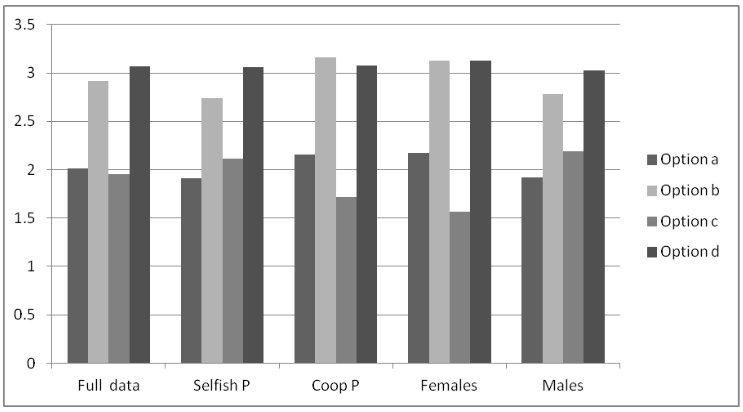

Fig. 2. Average rank by user group of responses to question 8

– The results for users with the more selfish personality values are very similar to the results for male users; and the results for users with the less selfish personality values are very similar to the results for female users. This follows from the background observation we made, that the male students had on average more selfish personalities than the female students.

The results for the eighth question are in Figure 2. In this Figure, if the bar for, e.g., option (a) is shorter than the bar for option (b), then this means that, on average, users gave option (a) greater importance than option (b).

Looking first at the results for the full set of users, we see that on average they ordered the options in decreasing importance as follows: good group experience (option c); good quality movie (option a); high rating (option b); and high satisfaction with the recommendation (option d). From the Figure, we see that the first two options are very close in their average rank. Bear in mind, though, that this experiment has more males than females and hence more users who, on average, are more selfish. A clearer picture emerges when we look at these different types of user separately.

If we look at users with less selfish personalities (and, equally, the female students in this experiment), we see that this ordering is accentuated: the group experience (option c) is more markedly important than the movie quality (option a), and there is more equivocation between options (b) and (d). But for users with more selfish personalities (and, equally, the male students), we see that the ordering of the first two options is reversed: recommending a good quality movie (option a) is more important than recommending a movie that results in a good group experience (option c). It is perhaps no surprise that more selfish users treat the group experience as less important. It is interesting though that movie quality is more important than whether they like the movie (option b) and whether they are satisfied with the recommendation (option d).

Overall, there are two surprises in the results. First, across all users the idea that a recommender does a good job when it recommends the movie that the users would have gone to see in reality (option d) is always treated as being of low importance. Second, across all users 'objective' movie quality is important: perhaps we need to ensure that we recommend items whose expert reviews or population average ratings exceed a minimum quality.

It would be unwise to draw firm conclusions from experiments like this one, particularly because the questions make rather subtle distinctions which the respondents may have misunderstood and the number of respondents is quite low. What we are probably safe to conclude is the importance of the group experience, the importance too of choosing high quality movies, and the sense that, if there is a trade-off to be made, the less selfish people are the ones who can remain satisfied even when the trade-off is at their expense.

6 Discussion

Our investigation has implications for the design of group recommender systems.

A first implication is that group recommender systems need to model, and hence predict, the three dimensions. For each candidate movie, they need to predict how much each user will like the movie; how satisfied the group members will be with the different ways in which their preferences are traded-off; and the group experience. Our experimental results suggest that it may even be important to be able to predict some sort of 'objective' movie quality, since this was given high importance by the students in the experiment.

One way a recommender can predict these factors is for us to *design* prediction models. Nearly all work on group recommender systems has taken this approach to the prediction of users' satisfaction with the recommendation. This is what the different aggregation functions do, including our own social recommender that takes personalities and trust into account (Section 2). But designing such models is difficult. There is a risk that our models are too simplistic, failing to take into account the richness of group dynamics.

A better approach might be to try to *learn* these models, using the feedback that we have been discussing to give us training data. This, after all, is how we predict single-user ratings. Why should we not take the same approach to predictions of recommendation satisfaction and of the group experience? An approach that generalises from training data might be more sensitive to nuances in the ways that groups operate. The case-based variant of our group recommender system (Section 2) works in this way —at least, in a simple-minded form: aggregation is based on 'replaying' the decision-making from similar movie-going events. It does not go so far as to predict the group experience.

CBR might be very well-suited to this task. After all, CBR is all about reasoning with experiences [1]. Since groups recur (with small variations) and groups structures (such as a parent and his or her children, or a group of university-age friends) recur, the CBR assumption (similar problems have similar solutions [8]) might apply. A rich case structure can capture multiple aspects of the movie-going event. The problem description part of the case can contain some or all of the following: (a) information about each member of the group —demographic information, personality information, and information about tastes, e.g. in the form of ratings; (b) information about relationships between group members; (c) the candidate movies, i.e. the ones from which the recommender made its recommendations; (d) predicted ratings for each group member and each candidate movie; and even (e) predictions about the other dimensions (user satisfaction and the group experience). The solution part of the case can contain at least the movie or movies that were recommended and might contain more than this (e.g. the ranking of all the candidate movies).

But to make good recommendations, we cannot simply retain cases of this kind in a case base and replay them. The case may be suboptimal; the movie that the group went to see may not have been the best movie for this group. If we retain it, we will replay it in any future recommendation where it gets retrieved as a neighbour, where it may contribute to suboptimal decisions in the future. We need to store information about how successful each case is. Cases can include a third component (alongside the problem description and the solution), namely the outcome [6]. In a recommender system, the outcome records user feedback —the main subject of this paper. The feedback can be compared with predicted values to give a measure of the (sub)optimality of the case.

But there remains a question of practicality. We suspect that users will be either unwilling or unable to give each of the three kinds of feedback. Furthermore, when current group recommender systems ask their users for a movie rating, it is probable that users do not wholly distinguish between movie ratings (whether they liked the movie), satisfaction with the recommendation (whether the recommender traded-off preferences in a good way) and the group experience. The movie rating they supply is likely to be influenced by the other two factors.[6]

Perhaps if group recommender systems are to ask for only one form of feedback, they should instead ask users for just their rating of the group experience. This is easily understood: “On a scale of 1 to 5 (where 1 means ‘Not at all’ and 5 means ‘A very great deal’), how much did you enjoy watching this movie with your friends?” This by no means solves all the problems we face in building a new generation of group recommender systems. If we ask for only one form of feedback, we then face a *credit assignment problem*: determining how much of their enjoyment (or lack of it) was attributable to various factors, and representing and reasoning with the uncertainty that arises from this credit assignment. Furthermore, in a group recommender, we may have varying degrees of feedback incompleteness: some group members may return to the system and supply a rating; others may not, and this increases uncertainty and introduces bias.

We cannot conclude this paper with a design prescription. But we hope that our reflection on our experience of building a number of group recommender systems, along with some of the insights that come from our experiment, suggest a direction of travel for future work or, at least, will provoke useful discussion.

[6] Ratings in single-user recommenders also exhibit contextual influences [7]. But, here we are focussing on issues that are specific to, or accentuated in, group recommender systems.

References

1. R. Bergmann. *Experience Management: Foundations, Development Methodology, and Internet-Based Applications.* Springer, 2002.
2. D. Bridge, M. H. Göker, L. McGinty, and B. Smyth. Case-based recommender systems. *Knowledge Engineering Review*, 20(3):315–320, 2005.
3. J. Golbeck. Generating predictive movie recommendations from trust in social networks. In *4th International Conference on Trust Management*, pages 93–104, 2006.
4. A. Jameson and B. Smyth. Recommendation to groups. In *The Adaptive Web, Methods and Strategies of Web Personalization*, pages 596–627. Springer, 2007.
5. B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *User Model. User-Adapt. Interact.*, 22(4-5):441–504, 2012.
6. J. Kolodner. *Case-Based Reasoning.* Morgan Kaufmann Publishers Inc., 1993.
7. Y. Koren and R. M. Bell. Advances in collaborative filtering. In *Recommender Systems Handbook*, pages 145–186. Springer, 2011.
8. D. B. Leake, editor. *Case-Based Reasoning: Experiences, Lessons, and Future Directions.* MIT Press, 1996.
9. J. Masthoff and A. Gatt. In pursuit of satisfaction and the prevention of embarrassment: affective state in group recommender systems. *User Modeling and User-Adapted Interaction*, 16(3-4):281–319, 2006.
10. L. Quijano-Sánchez, D. Bridge, B. Díaz-Agudo, and J. A. Recio-García. A case-based solution to the cold-start problem in group recommenders. In *20th ICCBR*, pages 342–356. Springer, 2012.
11. L. Quijano-Sánchez, J. A. Recio-García, B. Díaz-Agudo, and G. Jiménez-Díaz. Social factors in group recommender systems. *ACM Transactions on Intelligent Systems and Technology*, 4(1), 2013.
12. F. Ricci, B. Arslan, N. Mirzadeh, and A. Venturini. ITR: A Case-Based Travel Advisory System. In *6th ECCBR*, pages 613–641. Springer, 2002.
13. L. Quijano Sánchez, D. Bridge, B. Díaz-Agudo, and J. A. Recio-García. Case-Based Aggregation of Preferences for Group Recommenders. In *20th ICCBR*, pages 327–341. Springer, 2012.
14. L. Quijano Sánchez, J. A. Recio-García, and B. Díaz-Agudo. User satisfaction in long term group recommendations. In *19th ICCBR*, pages 211–225. Springer, 2011.
15. K.W. Thomas and R.H. Kilmann. *Thomas-Kilmann Conflict Mode Instrument.* Tuxedo, N.Y., 1974.