Observing User Behavior in Group Recommender Systems

Extended Abstract

keywords: recommender systems; social choice theory; user modeling

1 Overview

Recommender systems help users decide among a multitude of choices, and thus address the information overload present in our everyday lives. Research in the field received rekindled interest following the Netflix prize, where advanced *Collaborative Filtering* (CF) techniques demonstrated their superiority [7]. CF is based on the premise that the behavior of users in the past, in terms of their consumption and feedback, can help predict future behavior.

In many cases, a recommendation to a group of people rather than a single person is required. For example, consider some friends planning their summer vacation destination [1], or a family deciding on a TV program to watch [9]. The additional challenge in making group recommendations is how to combine individual preferences. For this problem, ideas from *Social Choice* theory have been employed. In particular and similar to how other computer science fields were influenced e.g., [5], various *aggregation strategies* of individual preferences have been proposed with the understanding that no single one can be optimal in a formal sense [2].

To design better group recommender systems, it is first important to improve our understanding of which aggregation strategy works better and in what setting. Towards this direction, two distinct paradigms have been followed. The first is by case studies, where the goal is to observe how people actually combine preferences and make group decisions [8, 4]. The second is to define evaluation metrics of aggregation strategies and then apply them over semi-synthetic data [3] (datasets with actual feedback from groups are scarce, often non-open). However, each paradigm has its own drawbacks. The former suffers in *generalizability* and cannot scale to the order of millions of users and thousands of items, where recommender systems typically apply. The latter suffers from *experimenter bias*, as the choices made in designing the evaluation setting tend to favor group recommenders with specific aggregation strategies.

Our first contribution directly addresses the aforementioned shortcomings. We design a technique to extrapolate the observations from small-scale studies to meaningful real-life scales. We try to match well-known aggregation strategies from social choice theory to the observed group decisions, and identify a mixture of strategies that appears to match well group behavior. We then apply these strategies over large publicly available datasets, typically used in recommender system research, to generate more realistic semi-synthetic group behavior. Our evaluation setting minimizes experimenter bias by considering multiple evaluation criteria.

Our second contribution is a novel CF machine learning technique that attempts to discover how a group behaves.[1] It does so by observing user behavior individually and within groups,

[1] In a pure CF setting, user and group behavior refers to explicit feedback (i.e., ratings) given to items.

Table 1: Suitability of aggregation strategies for describing user behavior in the study of [4]

	RMSE	max-RMSE	NDCG@5	MAP
AVERAGE	1.041	**1.814**	0.676	**0.218**
LEAST-MISERY	**0.919**	2.170	**0.682**	0.209
MAX-PLEASURE	1.766	2.518	**0.682**	**0.218**

and measuring the discrepancies. Then, these discrepancies are translated into *behavior roles* that the user assumes within a group, e.g., a leader within the group has stronger influence on the group decisions. Under our experimental setting, this approach is shown to be able to quickly learn user behavior, and have better predictive power compared to existing group recommenders with rigid preference aggregation strategies.

2 Results

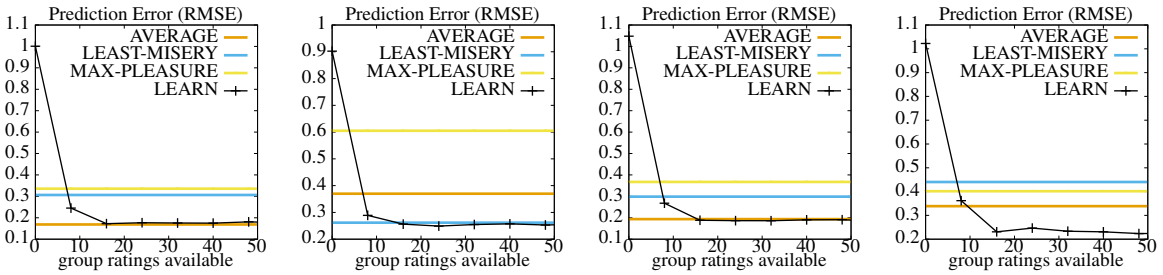
As a first step, we investigate the results from the observational study of [4]. Students from four universities were arranged into groups of 2–4 members. Each member was asked to individually rate on a 5-point scale the attractiveness of 11 popular European capitals as a touristic destination. Then, the groups convened and jointly agreed on their top-2 preferred destinations. Overall, there were 200 users partitioned across 60 groups.

We consider three popular aggregation strategies from the literature [8], which were adapted from social choice theory. Namely, AVERAGE assigns equal weight to the opinion (in our case, rating of a destination) of all group members and assigns the average opinion as the group opinion. LEAST-MISERY attempts to minimize the chance that any single member will be strongly dissatisfied with the group choice; hence, the least favorable opinion becomes the group opinion. Conversely, MAX-PLEASURE seeks to maximize the chance that any single member will be strongly satisfied with the group choice; thus, the group adopts the most favorable opinion among its members.

We wish to determine which strategy behaves best for each group. As observed in [4] and independently here, there is no clear answer; groups behave in different ways and cannot be described under a single aggregation strategy. Table 1 shows how the popular strategies behave under four standard performance criteria [6]; the bold value in a column represents the best performance for the particular criterion. The first two criteria measure prediction error, i.e., discrepancy between observed and predicted behavior, in terms of root-mean-square-error over all groups (RMSE) and for the worst-case group (max-RMSE); naturally lower values are better. The last two criteria measure how well the predicted ranking of destinations matches the actual ranking decided by the group, in terms of normalized discounted cumulative gain at rank 5 (NDCG@5) and mean average precision (MAP); higher values are desired. Each strategy scores two wins with none convincing.

Motivated by these findings, we then seek to construct realistic large semi-synthetic datasets to experiment upon. We start with the popular MovieLens 1M dataset that contains over one million ratings, on a scale of 1 to 5, by about 6000 users for about 4000 movies.[2] We then construct groups by assigning users to groups uniformly at random. To generate group ratings, we assume that groups behave according to exactly one of the three aforementioned aggregation strategies, or to a fourth one termed DICTATORSHIP, where a single arbitrary member within the group decides for the rest. While existing group recommender techniques can easily capture the behavior in each of the first three, they fail under DICTATORSHIP.

[2] http://grouplens.org/datasets/movielens/



(a) AVERAGE (b) LEAST-MISERY (c) MAX-PLEASURE (d) DICTATORSHIP

Figure 1: Prediction error of LEARN w.r.t. existing methods for various group strategies

Furthermore, we develop a CF-based technique, termed LEARN, that teaches itself how the group decides by comparing group to member behavior. We perform an experiment over the semi-synthetic data, where we measure the prediction error (RMSE) as we progressively increase the size of the observed group behavior, i.e., number of group ratings given; analogous results hold for max-RMSE, and ranking criteria. Results are shown in Figure 1. Note that existing methods AVERAGE, LEAST-MISERY, MAX-PLEASURE explicitly assume a fixed group behavior and cannot adapt according to observed group behavior; hence they are depicted as constant lines for reference. One should remember that each method has a favorable setting, the homonymous group behavior; e.g., LEAST-MISERY performs best in Figure 1(b) where groups behave under the least-misery principle. The important observation is that our LEARN method can successfully learn the behavior of each group just after seeing 10–15 group ratings and match or exceed the performance perform of the ideal method. Moreover, for the hard case of DICTATORSHIP where no ideal method exists, LEARN significantly outperforms them.

Overall, we draw the following conclusions: (1) there is no single aggregation strategy that best models group decision making; (2) in the absence of real data with observed group behavior at a large-scale, we can meaningfully extrapolate on small-scale case studies as long as we acknowledge the first point; (3) a variety of user behavior in groups can be quickly learned, at least on large-scale semi-synthetic data; (4) group and more general social recommender systems are prime vessels for testing social science and social psychology theories, but remain largely unexplored due to lack of publicly available large-scale data.

References

[1] L. Ardissono, A. Goy, G. Petrone, M. Segnan, and P. Torasso. Intrigue: Personalized recommendation of tourist attractions for desktop and hand held devices. *Applied Artificial Intelligence*, 17(8-9):687–714, 2003.

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

[3] L. Baltrunas, T. Makcinskas, and F. Ricci. Group recommendations with rank aggregation and collaborative filtering. In *RecSys*, pages 119–126, 2010.

[4] A. Delic, J. Neidhardt, T. N. Nguyen, F. Ricci, L. Rook, H. Werthner, and M. Zanker. Observing group decision making processes. In *RecSys*, pages 147–150, 2016.

[5] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW*, pages 613–622, 2001.

[6] A. Gunawardana and G. Shani. Evaluating recommender systems. In F. Ricci, L. Rokach, and B. Shapira, editors, *Recommender Systems Handbook*, pages 265–308. Springer, 2015.

[7] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, Aug. 2009.

[8] J. Masthoff. Group recommender systems: Aggregation, satisfaction and group attributes. In *Recommender Systems Handbook*, pages 743–776. 2015.

[9] Z. Yu, X. Zhou, Y. Hao, and J. Gu. TV program recommendation for multiple viewers based on user profile merging. *User Model. User-Adapt. Interact.*, 16(1):63–82, 2006.