User Perception of Differences in Recommender Algorithms

Michael D. Ekstrand[1,2], F. Maxwell Harper[2], Martijn C. Willemsen[3], and Joseph A. Konstan[2]

[1]Dept. of Computer Science
Texas State University
San Marcos, TX, USA
ekstrand@txstate.edu

[2]GroupLens Research
Dept. of Comp. Sci. & Eng.
University of Minnesota
Minneapolis, MN, USA
{harper,konstan}@cs.umn.edu

[3]Human-Technology Interaction Group
School of Innovation Sciences
Eindhoven University of Technology
Eindhoven, The Netherlands
M.C.Willemsen@tue.nl

ABSTRACT

Recent developments in user evaluation of recommender systems have brought forth powerful new tools for understanding what makes recommendations effective and useful. We apply these methods to understand how users evaluate recommendation lists for the purpose of selecting an algorithm for finding movies. This paper reports on an experiment in which we asked users to compare lists produced by three common collaborative filtering algorithms on the dimensions of novelty, diversity, accuracy, satisfaction, and degree of personalization, and to select a recommender that they would like to use in the future. We find that satisfaction is negatively dependent on novelty and positively dependent on diversity in this setting, and that satisfaction predicts the user's final selection. We also compare users' subjective perceptions of recommendation properties with objective measures of those same characteristics. To our knowledge, this is the first study that applies modern survey design and analysis techniques to a within-subjects, direct comparison study of recommender algorithms.

Categories and Subject Descriptors

H.1.2 [**User/machine systems**]: Human factors; H.3.3 [**Information storage and retrieval**]: Retrieval models

Keywords

recommender systems; user study

1. INTRODUCTION

The ability of a recommender system to meet the needs of its users depends on many aspects of the recommender's behavior, the application domain, and the user's information needs. We want to understand the relevant properties of each of these entities (recommenders, domains, and needs) and how they interact to form a compelling recommendation experience in a robust and systematic fashion. This leads to a key question: how do users perceive the outputs from various recommender algorithms to be different, and how those differences affect their choice of algorithm?

The research community has long known that there are subjective differences in the output of recommender algorithms, even among algorithms with comparable accuracy [21, 12]. To map out some of those differences in the movie recommendation domain, we present a user study aimed at understanding the subjective differences users perceive between different collaborative filtering algorithms and how those differences affect their choice of recommender system.

We are taking advantage of a unique opportunity: we have a large base of experienced recommender system users (the MovieLens user community) and a software toolkit capable of supporting a wide array of recommender algorithms (LensKit [5]). Further, we are preparing the general release of a new version of the MovieLens platform, providing an opportunity to conduct an experiment in a context where the question of user preference among recommender algorithms has real meaning. Finally, new insights from user evaluation studies allow us to measure the subjective aspects that explain *why* particular recommender algorithms are preferred [11].

This paper is one installment in a series of work on understanding how recommenders can best meet users' information needs. It builds on extensive work on offline recommender evaluation and previous results identifying factors that influence user preference among recommendations and recommendation lists, such as diversity and novelty. Follow-up work will need to examine a greater range of algorithms, contexts of use, and properties that mediate a recommender's ability to satisfy its users needs. Critically, it will also need to look at users' long-term satisfaction with their recommender choices.

In the present work we seek to answer the following questions:

RQ1 How are users' overall preferences for recommendation lists predicted by the subjective properties of those lists?

RQ2 What differences do users perceive between the lists of recommendations produced by commonly-used collaborative filtering algorithms?

RQ3 How do objective algorithm performance metrics relate to users' subjective perception of recommender outputs?

To that end, we present the results of a user study we conducted users of the MovieLens recommender system, asking them to compare recommendation lists produced by popular recommender algorithms. We specifically explore item-item, user-user, and SVD algorithms, looking at user perceptions of accuracy, personalization, diversity, novelty, and overall satisfaction. Each user provided a first-impression preference between a pair of algorithms, subjective comparisons of the algorithms' output on our dimensions of interest, and selected an algorithm for future use. We build a model that predicts both the user's initial preference and their final choice of algorithm the subjective perceptions and objective measures of the recommender algorithms and their output.

RecSys 14 Silicon Valley, CA USA

http://dx.doi.org/10.1145/2645710.2645737.

While this experiment focuses on one recommendation domain that is admittedly well-studied, it uncovers subjective characteristics of recommender behavior that explain users' selections in a manner that provides a good basis for generalization, replication, and further validation. We report specific relationships that can be tested for validity in additional contexts, providing much greater insight into what aspects of algorithm suitability for movie recommendation are task-specific and what are more general behaviors.

In addition to answering our immediate questions, the data collected in this survey should be a useful ground truth for calibrating new offline measures of recommender behavior to more accurately estimate how algorithms will be experienced by their users.

2. RELATED WORK

Many researchers have acknowledged the role of factors beyond accuracy — either predictive or retrieval accuracy — in evaluating recommender systems [13]. This has resulted in the development of offline evaluation protocols that incorporate metrics beyond accuracy [8, 20], user-based research on recommender perception [11, 16], and a number of workshops and tutorials on recommender system evaluation. Industrial applications often evaluate recommender approaches by measuring lift, click-through rates, and other observable user behaviors that affect the core business goals the recommender is intended to serve; these behaviors arise from the holistic impact of the recommender on the user's actions.

User studies are widely used to evaluate the usefulness of particular recommender applications [4] and to answer scientific questions about user interaction with recommender systems [1]. The design and execution of user studies has improved over time; historically, many studies involved relatively simple user questionaires (a practice that continues today), but recent years have seen increasing development and use of more sophisticated study designs and analysis techniques [11].

One such technique, structural equation modeling [10, 11], is a powerful tool for investigating the perceived factors that influence user satisfaction and choices. It allows us to not only measure what algorithms or items the user ultimately prefers, but also assess how specific aspects of the recommendations (such as novelty and diversity) influence their preferences and behavior. A user may prefer algorithm A over B because it is diverse and therefore more appropriate to meeting their needs, and SEM allows us to quantify and test these kinds of relationships.

The particular factors that we consider are motivated by a long line of work in human-recommender interaction and recommender user experience [14, 11, 16]. Novelty [25, 22] and diversity [27, 22, 26, 24] are both widely regarded as an important factor in recommender system perception and acceptance.

3. EXPERIMENT DESIGN

To assess the differences among various algorithms for recommending movies with explicit user ratings, we conducted an experiment in which users reviewed two lists of recommendations and took a survey comparing them. Figure 1 shows a screenshot of the experimental interface.

3.1 Users and Context

We conducted our experiment on users of MovieLens, a movie recommendation service. The survey was integrated into a beta launch of a new version of MovieLens; we invited active users to preview the beta with an on-site banner and required them to participate in the experiment prior to using MovieLens Beta. 1052 users attempted the survey, of which 582 completed it. Since we limited recruiting to active users, all users had at least 15 ratings (the median rating count was 473).

3.2 Algorithms

For this experiment, we tested three widely-used collaborative filtering algorithms as implemented in LensKit version 2.1-M2 [5]. To tune the algorithm parameters, we used the item-item CF configuration in the MovieLens production environment and values reported in the published literature [5, 6] as a starting point and refined the configurations with 5-fold cross-validation over the MovieLens database (using RMSE and prediction nDCG as our metrics to optimize) and manual inspection of recommender output. This resulted in the following algorithm configurations:

- Item-item CF [19] with 20 neighbors, model size of 4000, cosine similarity, item mean centering, neighbor threshold of 0.1, and requiring 2 neighbors to make a prediction.
- User-user CF [7] with 30 neighbors, cosine vector similarity between users, and normalizing user ratings by subtracting the personalized user-item mean, a neighbor threshold of 0.1, and requiring 2 neighbors to make a prediction; we additionally applied a small Bayesian damping of 5 to the user and item means for normalization.
- SVD with the FunkSVD [6, 15] training algorithm, using 50 features, 125 training epochs per feature, user-item mean baseline with damping of 5, and the LensKit default learning rate of 0.001 and regularization factor of 0.015.

For each user, we randomly selected two of the algorithms. For each algorithm, we computed a recommendation list containing the 10 movies with the highest predicted rating among those the user had not rated, sorted by predicted rating. We presented these lists as 'List A' and 'List B' (the ordering of algorithms was randomized).

In internal pre-testing, the user-user and SVD algorithms often recommended very obscure movies. This created a significant risk that users would be entirely unfamiliar with the recommendations of these algorithms. While we want to measure novelty, users are limited in their ability to judge completely unfamiliar lists. To test the algorithms in something close to their pure form, while increasing the likelihood that users would have at least heard of some of the movies and therefore be able to provide meaningful feedback, we limited each algorithm to recommending from the 2500 most-rated movies in MovieLens (about 10% of MovieLens's entire collection). This adjustment may limit effect sizes (e.g. decreasing perceived novelty of an algorithm's recommendations), but should allow each algorithm to still demonstrate its general behaviors relative to the others.

Not all algorithms could produce 10 recommendations for all users. If a user could not receive 10 recommendations from each algorithm, we exclude them from the analysis.

Most studies of recommender user experience employ a between-subjects design in which the users only see one condition (i.e. one algorithm at the time). Such between-subject designs are more realistic of real world experiences. However, in our present experiment we are primarily interested in detecting differences between algorithms, some of which may be quite subtle. If users evaluated each algorithm's output separately, their experience with that algorithm would not be related to another; this is problematic as evaluation is a naturally relative activity: absolute judgments are much more difficult than relative judgments and less sensitive to small differences [9]. Therefore we chose to evaluate these algorithms with a simultaneous within-subjects design in which our participants jointly evaluate two out of three algorithms side-by-side.

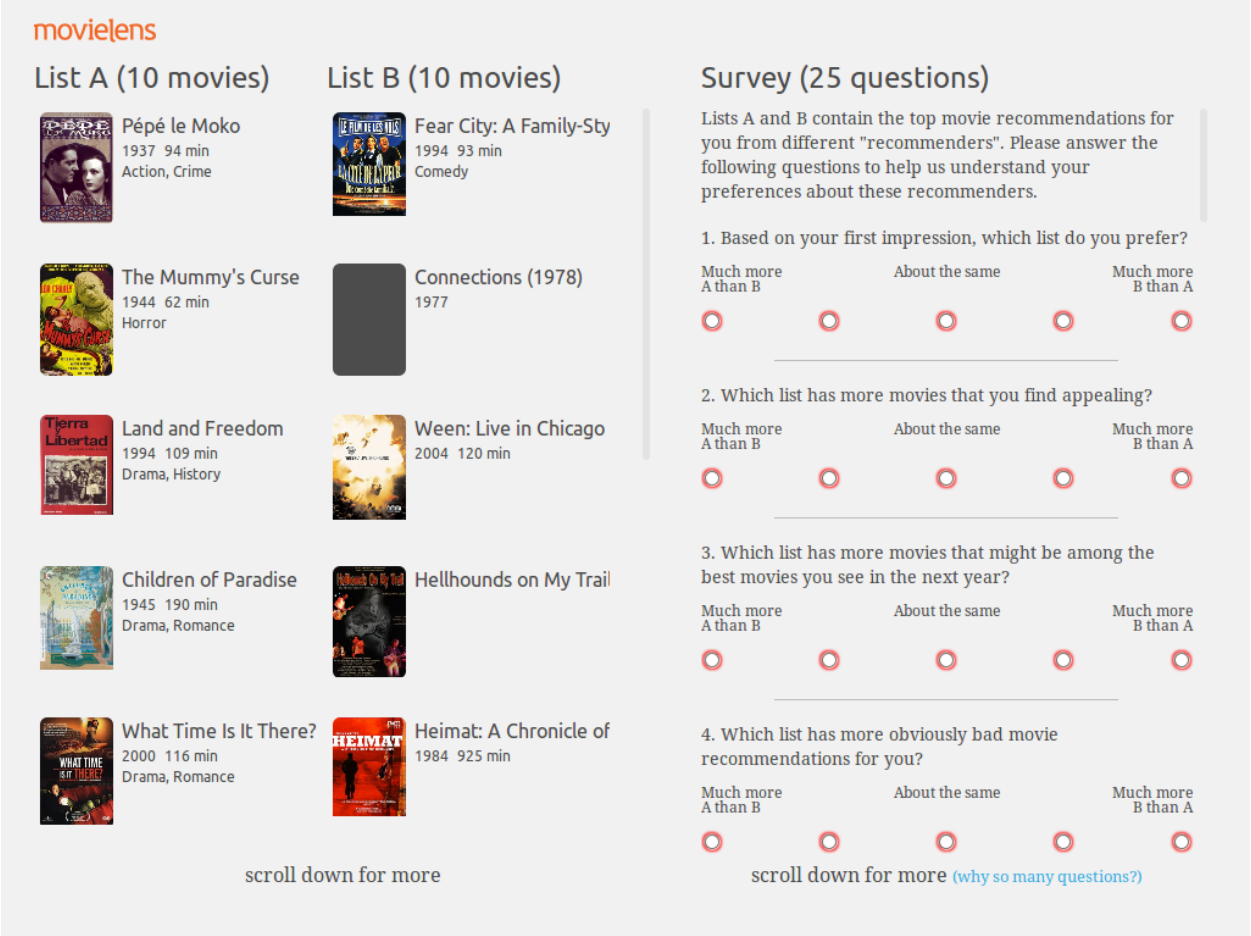


Figure 1: Screen shot of the experiment interface. Clicking on a movie in the list opens a pop-over with additional movie details.

3.3 Showing Predictions

Algorithms will not necessarily generate scores on the same portions of the rating scale. For example, one algorithm may tend to predict 4.5–5 stars, while another algorithm may be more conservative and predict 3.5–4.5 stars. While we want users to evaluate the recommendation lists, not the predictions, this could have a confounding affect if the predicted rating affects the user's perception of the movie lists. To control for this, we assigned each user randomly to one of the following prediction conditions:

- Show no predictions (just the list of recommended movies).
- Show a standard, unadjusted prediction.
- Show a normalized prediction. In this condition, we predicted the first 3 movies at 5 stars, the next 4 at 4.5, and the last 3 at 4 stars.

If predicted ratings do not affect the user's perception of the recommendation lists, then there should be no difference in response between these conditions and we can average across them in the final analysis.

3.4 User Survey

Our survey consists of four parts. The first question, visible in fig. 1, asks users which list of recommendations they prefer, based on their initial impression. 5 options are available, with the extremes labeled 'Much more A than B' and 'Much more B than A'.

Following initial preference are 22 questions about various aspects of the lists, designed to measure the user's perception of the recommendation lists across five factors:

Acc Accuracy — the recommender's ability to find 'good' movies.

Sat Satisfaction — the user's overall satisfaction with the recommender and their perception of its usefulness.

Und Perceived personalization ('Understands Me') — the user's perception that the recommender understands their tastes and can effectively adapt to them.

Nov Novelty — the propensity of the recommender to suggest items with which the user is unfamiliar.

Div Diversity — the diversity of the recommended items.

For this portion of the survey, we started with questions that have worked well in previous experiments [11], adapted them to be comparative instead of absolute, and wrote a small number of new questions. The full list of questions is in table 1a.

After the main body of questions, we ask users which algorithm they would like to use by default once MovieLens gains the ability to support multiple recommender algorithms in parallel (a feature we are planning to develop in the coming months). This question is forced-choice, requiring users to pick one of the two algorithms. It also carries some consequence for users: while they will be able to switch algorithms in their user settings page without much difficulty, the algorithm they select will be providing their default recommendations in the future.

3.5 Objective Metrics

In addition to soliciting users' subjective perceptions of the recommendations, we computed objective measures of the algorithms' behavior with respect to accuracy, novelty, and diversity.

We estimate the accuracy of each algorithm by computing the RMSE of using it to predict each user's last 5 ratings prior to taking the survey, averaging the errors per user. To estimate novelty, we take the simple approach of computing the mean popularity rank of the items recommended to the user (fig. 3b); this creates an 'obscurity' metric, where high values correspond to lists with more obscure items.

We compute diversity with intra-list similarity [27] using cosine between tag genome vectors [23] as the itemwise similarity function and normalizing the final metric so that a list of completely similar items has a score of 1; we exclude items for which tag genome data is not available (no list required us to exclude more than 2 items); fig. 3c shows these values.

To convert the metrics into comparative measures, we take the log ratio of the objective metric values for the two recommendation lists presented to a user[1]. This produces a single value for a pair of algorithms or recommendation lists that we can attempt to correlate with the users' subjective comparative judgements.

4. RESULTS

582 users completed the study over 81 days. Table 2 shows how many participated in each algorithm condition, along with their final choice of algorithm. Users generally selected both item-item and SVD over user-user ($p < 0.0001$), but there was no statistically significant difference in the proportion of users choosing between item-item vs. SVD. Table 1b summarizes the responses to each of our questions by algorithm condition, and fig. 3 shows the objective measures of each algorithm's output.

We observed no significant effect of either the ordering of algorithms or of the prediction condition[2], so we exclude those from the remainder of the analysis.

4.1 Response Model

To answer our more detailed research questions about the factors at play in users' choice of algorithms, we subjected the survey results to confirmatory factor analysis (CFA) and structure equation modeling (SEM). We used Lavaan [18] for R [17] to compute the CFA and SEM, treating all question responses as ordinal variables. Each question is mapped to the factor it was designed to target. Table 1c shows the question/factor loadings from both the initial CFA and a simplified SEM derived from it[3]. In the full CFA, there are several questions that have very low explanatory power (such as 'which recommender more represents mainstream tastes?' with $R^2 = 0.006$); in addition, the *Accuracy*, *Satisfaction*, and *Understands Me* factors are very highly correlated (corellation coefficents in excess of 0.9), so we cannot legitimately consider them to be measuring different constructs in this experiment. We simplify the model by removing the *Accuracy* and *Understands Me* factors (we retain *Satisfaction* because it has the highest explanatory power, as measured by the Average Variance Extracted, and all 5 of its questions load strongly), and removing poorly-loading questions from *Novelty*. We then expand the simplified CFA into an SEM, which we call the *Overall SEM*, by adding structural relationships between factors, regressing them against the experimental conditions and objective metrics, and regressing the user's first impression and final selection against the experimental factors.

Figure 2 and table 1c show the structure and question/factor loadings in this overall model. The overall SEM has good fit ($\chi^2_{139} = 229.5$, $p < 0.001$, CFI = 0.999, TLI = 0.998, RMSEA = 0.033). The model uses standardized factor scores, so a coefficient for the effect on or of a factor measures the effect in standard deviations of the factor. We use item-item vs. SVD as the baseline condition, encoding the item-item/user-user and SVD/user-user conditions with dummy condition variables.

4.2 RQ1: Predicting Preference

To address RQ1, we consider the impact of the factors (*Nov*, *Div*, and *Sat*) on the user's first impression of the recommendation lists and on their final choice of algorithm (see fig. 2). Most users who preferred one algorithm over the other at their first impression picked that algorithm in the final forced-choice question.

The only significant predictor (besides first impression) of the user's final choice of algorithm was their relative satisfaction with the two recommendation lists. Users tended to pick the algorithm with which they expressed more satisfaction.

Satisfaction in turn is influenced by the novelty (negatively) and diversity (positively) of the recommended items. Novelty also has a small positive impact on diversity, suggesting that there is an upside to novelty (as it correlating with more diverse lists, which correlates positively with satisfaction) but a strong downside (users don't like recommendation lists full of unfamiliar items).

In addition to its indirect effect through satisfaction, novelty had an additional negative influence on the user's first-impression preference. This means that novelty has a strong initial impact on user preference. However, after the user has made their first judgement, answered the more in-depth questions, and finally selected an algorithm, the direct impact of novelty went away and their final choice depended primarily on satisfaction. Novelty is still a significant negative influence, but it is mediated through satisfaction.

4.3 RQ2: Algorithm Performance

In RQ2, we want to understand how the algorithms themselves compare on relative satisfaction, diversity, novelty, and user preference as exhibited in their choice of algorithm. Table 2 summarizes the final choice performance of the three algorithms: as measured by users picking an algorithm for use, user-user clearly loses, and item-item and SVD are effectively tied.

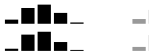
Table 1b provides some insight into users' perception of the relative charateristics of the algorithms. Across most questions, item-item and SVD are indistinguishable (user responses are symmetrically distributed about the neutral response). Item-item shows slightly more diversity than SVD. The other algorithm pairings show more differences across the board, with the exception of item-item and user-user being indistinguishable on diversity.

Our overall SEM (fig. 2) and related factor analysis incorporate the experimental condition, but its impact is difficult to interpret due to the comparative nature of the experiment. To better understand each pair of algorithm's relative performance, we reinterpret our experiment as three pseudo-experiments. Each of these pseudo-experiments uses one of the algorithms as a baseline and compares the other two algorithms on their performance and behavior relative to the baseline in a between-subjects design; the experimental treat-

[1] We also experimented with computing raw differences, but generally found the log ratio to be a better predictor.

[2] There are at least 3 possible causes of this non-effect: recommenders all predicted in the same range, prediction had no effect on perception, or our questions successfully guided users to evaluate the lists independent of prediction. In any case, it did not confound our results.

[3] Full Lavaan output for the SEM is included in the thesis form of this work [3].

Factor / Question (*W. l.* = 'Which list', *W. r.* = 'Which recommender')	II v. SVD	II v. UU	SVD v. UU	Full CFA Coef.	Full CFA R^2	SEM Coef.
First Impression						
Accuracy					**0.61**	
W. l. has more movies that you find appealing?				0.911	0.85	
W. l. has more movies that might be among the best movies you see in the next year?				0.786	0.64	
W. l. has more obviously bad movie recommendations for you?				−0.751	0.59	
W. r. does a better job of putting better movies at the top?				0.572	0.35	
Diversity					**0.64**	
W. l. has more movies that are similar to each other?				−0.772	0.61	−0.748
W. l. has amore varied selection of movies?				0.772	0.61	0.743
W. l. has movies that match a wider variety of moods?				0.838	0.71	0.806
W. l. would suit a broader set of tastes?				0.793	0.64	0.768
Understands Me					**0.63**	
W. r. better understands your taste in movies?				0.933	0.88	
W. r. would you trust more to provide you with recommendations?				0.943	0.90	
W. r. seems more personalized to your movie ratings?				0.842	0.73	
W. r. more represents mainstream tastes instead of your own?				−0.072	0.01	
Satisfaction					**0.82**	
W. r. would better help you find movies to watch?				0.923	0.86	0.737
W. r. would you be more likely to recommend to your friends?				0.846	0.73	0.678
W. l. of recommendations do you find more valuable?				0.884	0.79	0.717
W. r. would you rather have as an app on your mobile phone?				0.921	0.86	0.736
W. r. would better help to pick satisfactory movies?				0.928	0.87	0.745
Novelty					**0.43**	
W. l. has more movies you do not expect?				−0.770	0.64	0.750
W. l. has more movies that are familiar to you?				0.784	0.66	−0.762
W. l. has more pleasantly surprising movies?				0.454	0.24	
W. l. has more movies you would not have thought to consider?				−0.704	0.54	0.707
W. l. provides fewer new suggestions?				0.258	0.08	

(a) Questions

(b) Response distributions. Dark entries have significant bias (uncorrected Wilcox test, $p < 0.01$)

(c) CFA & SEM factor loadings.

Table 1: Overview of survey results. All SEM factor loadings are significant ($p < 0.01$); factor R^2 is AVE.

Condition (*A* v. *B*)	*N*	Pick *A*	Pick *B*	% Pick *B*	*p*
I-I v. U-U	201	144	57	28.4%	0.000
I-I v. SVD	198	101	97	49.0%	0.831
SVD v. U-U	183	136	47	25.7%	0.000

Table 2: Final algorithm selection by condition. *p*-values are for two-sided proportion tests, $H_0 : a/b = 0.5$.

Baseline	Tested	% Tested > Baseline	*p*
ItemItem	SVD	48.99	0.0000
	UserUser	28.36	
SVD	ItemItem	51.01	0.0000
	UserUser	25.68	
UserUser	ItemItem	71.64	0.6353
	SVD	74.32	

Table 3: Split experiment summary. *p*-values are testing the null hypothesis that the user picked the tested algorithm over the baseline the same proportion of the time.

ment is the choice of algorithm to compare against the baseline. We will refer to this algorithm as the *tested* algorithm.

Randomization ensures that the behavior characteristics of the baseline algorithm are likely to be evenly distributed between the two sets of users encountering that algorithm, so we can (with some limitations) interpret relative measurements of one algorithm's comparison with the baseline as absolute measurements of that algorithm's behavior for the purposes of comparing with measurements of another algorithm against the same baseline.

Table 3 shows the pseudo-experiments, their conditions, and user's selections under this interpretation. The first pair of rows describes one of the three pseudo-experiments. Examining all users assigned to one of the two conditions involving item-item CF, we use item-item as the baseline and ask how often users picked user-user or SVD over the baseline. We can apply this interpretation to all questions and factors, not just selection. This allows us to make cleaner inferences at the expense of some statistical power.

For each experiment, we re-analyzed the data using SEM and basic regressions to predict the user's relative preference and final choice. Each SEM reused the factor loadings from the overall SEM but re-learned the relationships between factors, choice, and condition. We also omitted the objective metrics from these SEMs in order to focus on the subjective differences between the algorithms. The model structure for each experiment is a simplification of fig. 2.

In addition to the condition, we also consider the number of ratings in the user's history prior to joining the experiment as a proxy for their level of experience. It is possible for algorithms to perform differently for different users, or for more experienced users to judge recommendation lists differently. We computed the median number of ratings for the users participating in the experiment and set a condition variable indicating whether a particular had 'many' or 'few' ratings relative to the median.

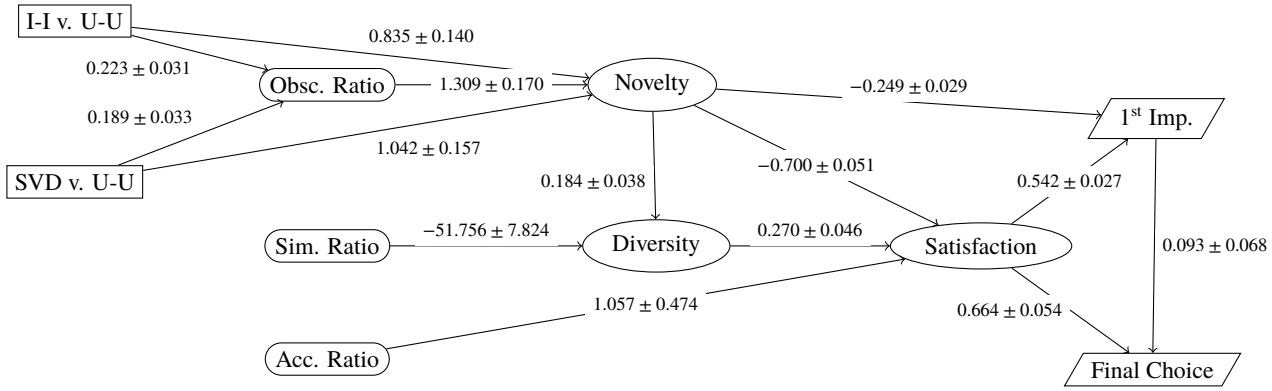
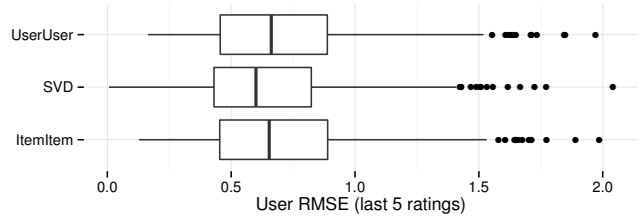
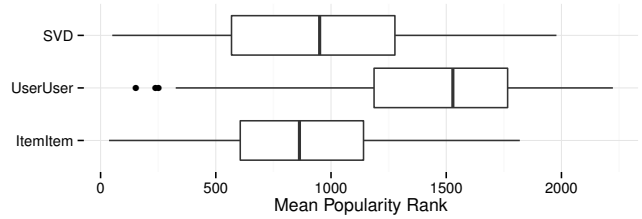


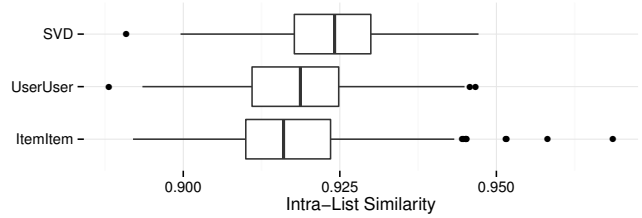
Figure 2: Overall SEM with bootstrapped standard errors. All displayed coefficients are significantly nonzero ($p < 0.01$). The baseline condition is I-I v. SVD; positive values & coefficients favor the right-hand algorithm (SVD or U-U).



(a) Algorithm predictive accuracy.



(b) Popularity rank of recommended items (lower is more popular).



(c) List diversity (lower is more diverse).

Figure 3: Objective recommendation list properties.

4.3.1 SVD vs. User-User

Users perceived user-user's recommendations to be more novel than SVD's (coef. 0.953, $p < 0.001$). They also reported user-user to be producing more diverse recommendation lists (coef. 0.312, $p < 0.001$). The effect on novelty was substantially stronger; combined with novelty's strong negative influence on preference, impression, and choice, users generally found SVD's recommendations more satisfactory and desirable than user-users. The effect of novelty on diversity was not present in this model; novelty only affected satisfaction directly.

As explained above, these results are from comparative judgements between the output of the tested algorithm (SVD or user-user) and the baseline algorithm (item-item). However, due to randomization, we assume that there are no important differences in item-item's output between the users comparing it against SVD and those comparing it against user-user. Therefore, we can reasonably make inferences about the relative behavior of SVD and user-user. These results are consistent with the raw survey response data for direct comparisons between SVD and user-user (table 1b), providing further support for their validity. They are also consistent with the objective measures of obscurity and diversity (figs. 3b and 3c).

Users selected SVD significantly more often than user-user (table 3), consistent with the results from users directly comparing SVD and user-user (table 2).

4.3.2 Item-Item vs. User-User

We found no significant difference in diversity between item-item and user-user CF; this is consistent with the raw results of direct comparison of these two algorithms in table 1b.

User-user produced more novel recommendation lists than item-item (coef. -1.563, $p < 0.001$). This effect interacted with user experience (rating count); for high-rating users, user-user's recommendations were not as novel as they were for low-rating users. This moderating effect was small, however, and user-user was significantly more novel than item-item even for high-rating users. There was no significant difference in item-item's novelty performance between low- and high-rating users.

4.3.3 Item-Item vs. SVD

Item-item produced slightly more diverse recommendations than SVD (coef. -0.26, $p < 0.001$); this is consistent with the response distributions in table 1b as well as the difference in intra-list similarity (fig. 3c). However, diversity did not have a significant influence on satisfaction in this pseudo-experiment: the only significant predictor of satisfaction was novelty.

The number of ratings the user had in their history prior to the experiment had a significant effect on the algorithm: for high-rating users, both algorithms were more novel than user-user. Since item-item and SVD did not have significantly different perceived novelty, this effect is reflecting user-user's decreased novelty for high-rating users. Whether there is an additional increase the novelty of item-item and SVD for high-rating users, or just a decrease in user-user's novelty, is beyond this experiment's capability to measure.

4.4 RQ3: Objective Metrics

To address RQ3, we consider in more detail the relationships between the objective metrics and subjective factors. Figure 3 shows the distributions of all objective metrics we computed.

The raw distributions of novelty and diversity measurements are consistent with the user survey results. User-user produces lists with less popular (and therefore likely more novel) items than SVD or item-item. SVD tends to produce somewhat less diverse recommendation lists. All three algorithms had comparable retrospective

accuracy, with SVD having a slight edge. Popularity/obscurity was the only objective metric that we found to significantly differ between conditions in the overall model.

Each objective metric was a statistically significant predictor of its corresponding subjective factor (fig. 2) and no other factor. Therefore, there is good correspondence between the subjective and objective measures of these three concepts, and the effect of the objective measures on final choice is completely mediated by their impact on the subjective measures. All indirect effects of objective measures on final choice are significant.

This means that predictive accuracy, for example, does affect the user's final choice, but only through the increased satisfaction that it produces. Further, the impact of novelty and diversity on satisfaction means that after controlling for predictive accuracy, diversity and novelty still have significant impacts on user satisfaction.

The direct effects of condition on novelty, in addition to the effect mediated through objective obscurity, suggest that user-user is producing lists that users perceive to be more novel beyond the sense of novelty that our objective metric can capture.

5. DISCUSSION

We set out to measure user perception of various interesting properties of the output of different recommender systems in a widely-studied domain. Our experiment uncovered mediation effects of novelty, diversity, satisfaction on users' choice of recommender algorithms. In this section, we highlight some of the key findings.

5.1 Effect of Novelty

One of the most striking things we found is that the novelty of recommended items has a signficiant negative impact on users' perception of a recommender's ability to satisfactorially meet their information needs. This effect was particularly strong in its impact on the user's first impression of an algorithm, and was present even though we restricted the depth of the long tail into which our algorithms could reach.

This suggests that recommender system designers should carefully watch the novelty of their system's recommendations, particularly for new users. Too many unfamiliar recommendations may give users a poor impression of a particular recommender, potentially driving them to use other systems instead. Increasing the novelty of recommendations as the user gains more experience with the system and has had more time to consider its ability to meet their needs may provide benefit, but our results cannot confirm or deny this. The users in our study are experienced with movie recommendation in general and MovieLens in particular (the median user has rated 473 movies), and their first impressions were still heavily influenced by novelty.

Our results complement the notion that that *trust-building* is an important goal of a recommender in the early stage of its relationship with its users [14]. They are also consistent with previous results finding that novelty is not necessarily positively correlated with user satisfaction or adoption of recommendations [2].

5.2 Diversity

We have also demonstrated that the diversity of recommendations has a positive influence on user choice of systems for general-purpose movie recommendation. Diversity is often framed as being in tension with accuracy, so that accuracy must be sacrificed in order to obtain diverse recommendation lists [27, 25, 26], and many diversification techniques do result in reduced accuracy by traditional objective measures. The strong correlation of perceived accuracy and satisfaction in our results provide evidence that there may not be such a tradeoff when considering user perception instead of traditional accuracy metrics.

The influence of novelty and diversity on satisfaction even after controlling for predictive accuracy provides direct, quantitative evidence for subjective but observable characteristics of recommendation lists affect user satisfaction and choice. Further, accuracy alone does not all aspects of satisfaction.

5.3 Algorithm Performance

When it comes to comparing the particular algorithms that we tested, item-item and SVD performed very similar, with users preferring them in roughly equal measure. We do not yet have insight into whether there are identifiable circumstances in which one is preferable over the other. It may be that one works better for some users than others; it may also be that their performance is roughly equivalent, and one does not work significantly better. The difference in the diversity of SVD and item-item, however, provides evidence for some kind of interesting difference between them.

User-user is the clear loser in our tests. Its prective accuracy was comparable to that of the other algorithms, but it had a significant propensity for novel recommendations that hurt both users' expressed satisfaction with its output and their interest in using it in the future. The lack of a significant independent effect of user-user condition on satisfaction or selection suggests that the increased novelty is the primary cause of user-user's poor subjective performance.

Finally, all three algorithms had similar predictive accuracy, but users still had strong preferences between some pairings. However, users selected item-item and SVD in almost equal numbers even though SVD had slightly higher predictive accuracy. This provides additional evidence that, at least beyond a certain point, offline metrics fail to capture much of what will impact the user's experience with a recommender system.

6. CONCLUSION AND FUTURE WORK

We have reported the results of a user study to compare the output of three common collaborative filtering algorithms and identify subjective, user-perceptible differences between them. This work directly advances our understanding of the role of diversity and novelty in how users evaluate recommender systems for potential use. We hope that the collected data will also be useful for developing and refining additional measures of recommender behavior, allowing for high-throughput offline evaluation to more accurately estimate the user experience with recommendations.

In the future, we plan at least two direct extensions of this work. First, we will examine users' long-term use of the algorithm switching feature we are developing for MovieLens; in particular, we want to see if users' expressed preference in our study corresponds to their long-term stable choice of algorithm. Second, we want to see if there are aspects of a user's profile beyond their experience level that predict their algorithm preference. Even though user-user did poorly overall, about 25% of users preferred it: who are these users, and why does it work for them?

We also want to explore additional objective measures that may predict the subjective characteristics we describe here.

Our results are currently limited to a single domain and task, although they are consistent with results elsewhere [24]. They are also limited to single configurations of each of the tested algorithms; alternative tunings may result in very different performance. The structural model and mediating factors we have described, however, provide a valuable starting point for understanding exactly how our results do or do not generalize. Further experiments in other domains can examine the subjective characteristics we have studied to see whether their relationships hold across recommendation tasks.

Understanding how users perceive and interact with recommenders is critical to building better tools for meeting users' information needs. This work provides new insights into the factors at work in the usefulness of movie recommendations. We look forward to much more work, from ourselves and others, to build out a more systematic understanding of how to produce effective, useful, and even delightful recommendations in a broad range of applications.

Acknowledgements

We thank our colleagues in GroupLens Research for their support and assistance. This work has been funded by the National Science Foundation under grants IIS 08-08692 and 10-17697.

References

[1] D. Bollen, B. Knijnenburg, M. Willemsen, and M. Graus. Understanding Choice Overload in Recommender Systems. In *Proc. ACM RecSys 2010*. ACM, 2010, pp. 63–70. DOI: 10.1145/1864708.1864724.

[2] Ò. Celma and P. Herrera. A New Approach to Evaluating Novel Recommendations. In *Proc. ACM RecSys 2008*. ACM, 2008, pp. 179–186. DOI: 10.1145/1454008.1454038.

[3] M. D. Ekstrand. Towards Recommender Engineering: Tools and Experiments in Recommender Differences. PhD thesis. Minneapolis, MN: University of Minnesota, 2014. 263 pp.

[4] M. Ekstrand, P. Kannan, J. Stemper, J. Butler, J. A. Konstan, and J. Riedl. Automatically Building Research Reading Lists. In *Proc. ACM RecSys 2010*. ACM, 2010, pp. 159–166. DOI: 10.1145/1864708.1864740.

[5] M. Ekstrand, M. Ludwig, J. A. Konstan, and J. Riedl. Rethinking the Recommender Research Ecosystem: Reproducibility, Openness, and LensKit. In *Proc. ACM RecSys 2011*. ACM, 2011, pp. 133–140. DOI: 10.1145/2043932.2043958.

[6] S. Funk. Netflix Update: Try This at Home. The Evolution of Cybernetics. 2006. URL: http://sifter.org/~simon/journal/20061211.html (visited on 04/08/2010).

[7] J. Herlocker, J. A. Konstan, and J. Riedl. An Empirical Analysis of Design Choices in Neighborhood-Based Collaborative Filtering Algorithms. *Inf. Retr.* 5(4):287–310, 2002. DOI: 10.1023/A:1020443909834.

[8] J. Herlocker, J. A. Konstan, L. Terveen, and J. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.* 22(1):5–53, 2004. DOI: 10.1145/963770.963772.

[9] C. K. Hsee and J. Zhang. General Evaluability Theory. *Persp. Psych. Sci.* 5(4):343–355, 2010. DOI: 10.1177/1745691610374586.

[10] R. B. Kline. *Principles and Practice of Structural Equation Modeling*. Guilford Press, New York, 1998. 354 pp.

[11] B. Knijnenburg, M. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the User Experience of Recommender Systems. *UMUAI*, 22(4):441–504, 2012. DOI: 10.1007/s11257-011-9118-4.

[12] S. McNee, N. Kapoor, and J. A. Konstan. Don't Look Stupid: Avoiding Pitfalls When Recommending Research Papers. In *Proc. ACM CSCW 2006*. CSCW '06. ACM, Banff, Alberta, Canada, 2006, p. 171. DOI: 10.1145/1180875.1180903.

[13] S. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *Ext. Abs. ACM CHI 2010*. ACM, Montréal, Québec, Canada, 2006, pp. 1097–1101. DOI: 10.1145/1125451.1125659.

[14] S. McNee, J. Riedl, and J. A. Konstan. Making recommendations better: an analytic model for human-recommender interaction. In *Ext. Abs. ACM CHI 2006*. ACM, 2006, pp. 1103–1108. DOI: 10.1145/1125451.1125660.

[15] A. Paterek. Improving regularized singular value decomposition for collaborative filtering. In *KDD Cup and Workshop 2007*, 2007.

[16] P. Pu, L. Chen, and R. Hu. A user-centric evaluation framework for recommender systems. In *Proc. ACM RecSys 2011*. In RecSys '11. ACM, 2011, pp. 157–164. DOI: 10.1145/2043932.2043962.

[17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[18] Y. Rosseel. lavaan: An R Package for Structural Equation Modeling. *J. Stat. Soft.* 48(2):1–36, 2012.

[19] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *Proc. ACM WWW '01*. ACM, 2001, pp. 285–295. DOI: 10.1145/371920.372071.

[20] G. Shani and A. Gunawardana. Evaluating Recommendation Systems. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pp. 257–297. Springer, 2010.

[21] R. Torres, S. McNee, M. Abel, J. A. Konstan, and J. Riedl. Enhancing Digital Libraries with TechLens+. In *Proc. ACM IEEE JCDL 2004*. ACM, 2004, pp. 228–236. DOI: 10.1145/996350.996402.

[22] S. Vargas and P. Castells. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proc. ACM RecSys 2011*. ACM, 2011, pp. 109–116. DOI: 10.1145/2043932.2043955.

[23] J. Vig, S. Sen, and J. Riedl. The Tag Genome: Encoding Community Knowledge to Support Novel Interaction. *ACM Trans. Interact. Intell. Syst.* 2(3):13:1–13:44, 2012. DOI: 10.1145/2362394.2362395.

[24] M. C. Willemsen, M. P. Graus, and B. P. Knijnenburg. Understanding the Role of Latent Feature Diversification on Choice Difficulty and Satisfaction. *Under review*, 2014.

[25] M. Zhang and N. Hurley. Avoiding Monotony: Improving the Diversity of Recommendation Lists. In *Proc. ACM RecSys 2008*. ACM, 2008, pp. 123–130. DOI: 10.1145/1454008.1454030.

[26] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *PNAS*, 107(10):4511–4515, 2010. DOI: 10.1073/pnas.1000488107.

[27] C.-N. Ziegler, S. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proc. ACM WWW 2005*. ACM, 2005, pp. 22–32. DOI: 10.1145/1060745.1060754.