Beyond accuracy: understanding user perception of diversity and serendipity in online movie recommenders

Avinash Akella[1], Joseph A. Konstan[2], Ruixuan Sun[2]

Department of Computer Science & Engineering[1], GroupLens Research[2]

Motivation

Recommender systems tend to show items that are similar to what a user has liked before. While accurate, these may be uninteresting to the user as they probably would have discovered them on their own anyway.

So recommender systems research has focused on beyond-accuracy metrics such as Serendipity, Novelty, Diversity, and Unexpectedness among others. These metrics are quite subjective and a lot of overlap exists in their definitions.

These have been shown to improve "user satisfaction", but how much of it aligns with actual user perceptions of them? Do users see them the same way? How much of it is based on user feedback? And are we meeting user expectations, and solving their problems?

We're trying to understand user perceptions of, and preferences for, different recommender objectives and the contexts that influence these decisions.

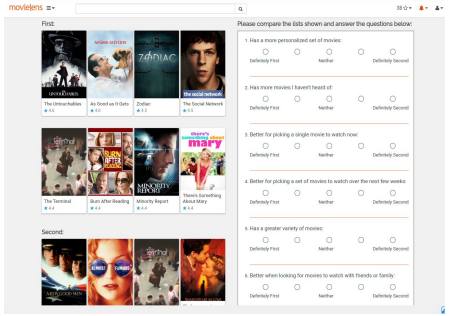
Objective

RQ1: Can users perceive these objectives as being different?

RQ2: In what contexts do they find them useful?

RQ3: How much does user preference or exploration behavior affect their choice?

Study - Design



Study - Recommenders

All are greedy re-rankers based on ITEM-ITEM collaborative filtering.

Baseline (T): ITEM-ITEM with popularity mix.

Diversity-improving (D): $f_{div}(i,R) = w_d.\text{rel}(i) + (1-w_d).\frac{1}{|C|}\sum_{j \in C} \text{dist}(i,j)$

Has more variety. Prioritizes the ones "farthest" from the rest of candidates.

Serendipity-improving (S): $f_{ser}(i,R) = w_s.\text{rel}(i) + (1-w_s).\min_{j \in C} \text{dist}(i,j)$

Has movies different from what the user has rated before. Picks movies "farthest" from movies in the user's rating history.

Study - Questions

The **List comparison questions** are designed to help us understand user perceptions of the algorithms on three fronts:

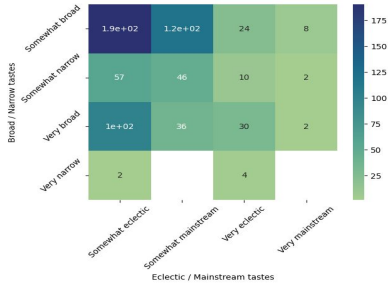
1. ***Metrics (M)***: Accuracy, Novelty, Diversity, Popularity, and Serendipity.
2. ***Context (C)***: Short-term goals, Long-term goals, and Group-watch.
3. ***Preferences (P)***: Dislikes, Overall Preference.

The **User Preferences questions** ask for a user's general movie-watching habits, namely:

1. How many movies do they watch a month? (***watch frequency***)
2. Do they have ***mainstream / eclectic*** tastes?
3. Do they like ***broad / narrow*** content?
4. Understand their priorities on MovieLens for:
   - Finding movies they've never heard of (***Novelty***)
   - Finding movies they heard of but forgot (***Unexpectedness***)
   - Finding movies similar to what they watched before (***Similarity***)
   - Finding a new variety of movies (***Diversity***)

Results

What's the right mix of user tastes?



Results

How were the recommenders perceived?

Question purpose	*Ranking*
Short-term	T >>>> D >>>>>> S
Long-term	T >>> D >>>>> S
Group watch	T > D >>>>>> S
Novelty	S >>>>>>>>> D > T
Serendipity	T >> S > D
Diversity	T > D >>> S
Dislike	S > D >>> T
Popularity	T > D >>>>>>>>>> S
Accuracy	T >>>> D >>>>> S
Overall Preference	T >>>> D >>>> S

What did our users have to say?

On Usage:

"...the **predictive algorithm doesn't always get it right**, but frankly with the tags, user rating, predicted rating, and 'More like this' **options all used together, it's SUPER easy to find movies** I enjoy".

"use it to compare my predicted score to the one of people I know".

On negative feedback:

"**filter out** films with specific actors".

"**be able to select that I don't like** action movies, or action scenes".

On group watch:

"**need a broader recommendation algorithm** that can recommend of several different subsets and genres rather than trying to match it all up as if it's for just one person".

On List comparisons:

" (**T**) **great for cozy watching** when you want something that is definitely good. (**D** has) things I have never heard of but seems intriguing, **good for long-term planning**.. or when you want to **be challenged by something you've never heard of** right now, and it seems to have a bigger focus on accessible family content. I **want a mix of both algorithms** on the page. Both great at different things".

"(**S**) has a mix of things I like and would probably like".

"(**S**) is **suddenly full of cheesy junk that I wouldn't watch"**.

Results

How did different user groups perceive them?

1. **Diverse users feel S is more serendipitous**, Non-diverse users feel it is T (for S vs T)
2. **Frequent movie-watchers have "broader" goals, and feel D is more accurate**. Infrequent users feel it is T. (for D vs T)
3. **Similarity-seeking users preferred D**, while non similarity-seeking preferred S. (for S vs D).
4. Perceptions for all the other groups: novelty-seeking vs non-novelty-seeking, eclectic vs mainstream users, fewer ratings vs more ratings, and expected goal vs unexpected goal users weren't very different.
5. Although, **perceptions sometimes differed when S was involved**.

Conclusion

1. Users seem to perceive novelty, popularity, and diversity clearly.
2. Diversity can be much more nuanced and tougher to generalize. Users can perceive diversity based on genres, release years, cast, tags etc.
3. Even users with niche tastes look for variety.
4. While watching with their friends and family, the expectation for personalization is lower, and it is higher for variety.
5. Lack of ways to incorporate negative feedback hurts both users, and the algorithm.

Future Work

1. Since perceptions for different user groups vary, how do we gather relevant information about users to understand and serve them better?
2. Effective controls can help us with this. But how do we design controls that are simple, intuitive, and clear to the user? That are useful and transparent?
3. Serendipity could be slight deviations from current tastes. But how much is "slight"? And along which direction do we deviate?
4. Every application has a different user-base. So every app needs its own analysis and its unique set of controls. It would be interesting to see how different user groups have different needs, and how we could adapt strategies to serve them.