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

AVINASH AKELLA, Department of Computer Science Engineering, United States
JOSEPH A. KONSTAN, Grouplens Research, University of Minnesota, United States
RUIXUAN SUN, Grouplens Research, University of Minnesota, United States

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.

CCS Concepts: • **Human-centered computing → Heuristic evaluations**; • **Information systems → Recommender systems**.

Additional Key Words and Phrases: diversity, serendipity, user perception

ACM Reference Format:
Avinash Akella, Joseph A. Konstan, and Ruixuan Sun. 2023. Beyond accuracy: understanding user perception of diversity and serendipity in online movie recommenders. *Proc. ACM Meas. Anal. Comput. Syst.* 37, 4, Article 111 (May 2023), 16 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recommender systems tend to show items that are similar to what a user has liked before [19]. While accurate, these may be uninteresting to the user as they probably would have discovered them on their own anyways [20]. For that reason, recommender systems research has focused on beyond-accuracy metrics such as Serendipity, Novelty, Diversity, and Unexpectedness among others. These metrics are quite subjective, and multiple definitions have been proposed to realize them [14].

Understandably, a lot of overlap exists in these definitions, especially between novelty, and serendipity [1, 19, 30]. For instance, Herlocker et al. [10] proposed that a serendipitous item has to be both novel and relevant. Adamopoulos and Tuzhilin [1], on the other hand, defined a new metric called 'unexpectedness' and argued that a serendipitous recommendation doesn't have to be novel as it can also be non-obvious or unexpected. By the latter argument, in the context of movie recommendations, a serendipitous item can be a movie that a user knows and would like to watch but forgot about because they haven't thought of it in a while.

Authors' addresses: Avinash Akella, Department of Computer Science Engineering, 200 Union Street SE, Minneapolis, Minnesota, United States; Joseph A. Konstan, Grouplens Research, University of Minnesota, 5-244 Keller Hall, 200 Union Street SE, Minneapolis, Minnesota, United States; Ruixuan Sun, Grouplens Research, University of Minnesota, 5-244 Keller Hall, 200 Union Street SE, Minneapolis, Minnesota, United States.

Many of these studies that try to improve on a defined metric have been conducted offline on real datasets [1, 13, 21, 23], while a few have conducted online analyses as well [4, 33]. A commonly reported metric in these online studies is user satisfaction which is measured either directly through a survey [4, 33], or indirectly through user activities like app acquisition (in Google Play)[5], or click-through rates [7]. 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? Some studies have tried to gather user perceptions on recommendations [3, 4, 33], but their goal is usually centered around evaluating or comparing a set of recommenders rather than trying to identify the gaps in user perception.

1.1 Motivation

Since much of the existing literature focuses on a single metric or objective function, there does not seem to be a proper consensus among them, and it isn't clear how they relate to each other [14]. We did not find an existing study that lets users interact with multiple algorithms: namely serendipity-improving, and diversity-improving simultaneously, and attempt to understand user perceptions of them.

In this work, we're trying to understand user perceptions of, and preferences for, different recommender objectives and the contexts that influence these decisions. Specifically, we're interested in answering the following questions through our user study:

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?*

We show users recommendation lists from two algorithms at a time, asking them to do pair-wise evaluations and see which one meets the criteria better. In Section 4.3.3 we try to address RQ1 by comparing algorithms that form a good representation of the serendipity-improving and the diversity-improving objectives. Specifically, we analyze user responses to see if they perceived the algorithms differently based on different criteria. In Sections 4.4.1, 4.4.2 we tackle RQ2, by analyzing user responses to context-based questions, and understanding how different metrics interact. Finally, in Section 4.4.3 we address RQ3, by comparing the perceptions of different user groups formed based on their self-declared, and inferred preferences. The scale of our study is sizeable with 600+ users completing the study, out of 3000+ enrollments. Hopefully, this can inspire future work on designing better experiences for human-in-the-loop systems.

2 RELATED WORK

2.1 History of user perception evaluation in recommender systems

Ziegler et al. [33] evaluated their greedy diversity-based re-ranking algorithm on top of a user-user or item-item Collaborative Filtering (CF) recommender, by asking users to rate the recommended lists based on diversity, relevance, and satisfaction. They found that light diversification in an item-based CF algorithm favors user satisfaction. Kotkov et al. [17] had users label serendipitous items to gather the first dataset with real user evaluations on the serendipity of items. They found that serendipity helps broaden user preferences but did not find any effects on user satisfaction. Celma [3] tried to evaluate user perceptions of relevance and novelty by having users compare recommendations from: an item-based collaborative model, a content-based model, and a hybrid of the two. They found that the content-based approach generated the most novelty and least accuracy, while the collaborative approach had the opposite effect.

They hypothesized that explanations could help improve the accuracy of novel items. Ekstrand et al. [6] asked users to compare pairs of recommendation lists to evaluate them based on perceived accuracy, novelty, and diversity. But they used an SVD factor model, a user-user, and an item-item CF model, none of which were serendipity-improving or diversity-improving by design. Willemsen et al. [29] tried to study the effects of levels of diversification by having 97 users compare three lists with increasing levels of diversity. They found that diversification beyond a certain level may not be appreciated by users, as we find out as well in our pilot study (see Section 3.3.4). While more studies exist on user perceptions of diversity [8, 9], some do look into user perception of serendipity as well. Zhang et al. [31] conducted a small-scale study to evaluate a serendipity-enhancing algorithm for movie recommendations with 21 participants. They had them compare recommendations from a baseline recommender with that from a serendipity-enhancing one and found that users prefer the less-accurate serendipity-enhancing system for discovering new artists. Chen et al. [4] conducted a large-scale user survey to gather feedback on serendipity, diversity, unexpectedness, timeliness, and relevance of e-commerce recommendations shown using three recommender algorithms optimized for serendipity, novelty, and relevance respectively. They found that Serendipity was more effective in improving user satisfaction, compared to diversity and novelty. But, unlike our study, users were divided into groups and each group was shown recommendations from only one algorithm.

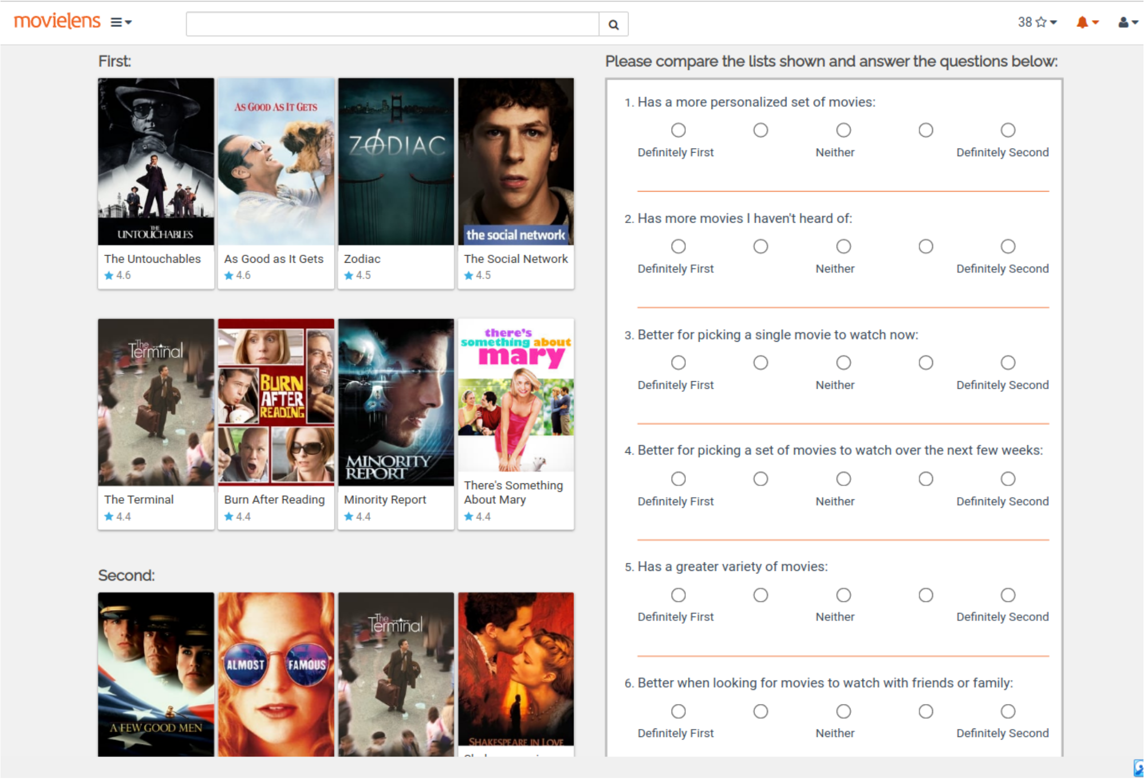
2.2 Serendipity-improving recommenders

Iaquinta et al. [12] were among the first to suggest serendipity-improving recommender systems. They employed a content-based recommender system that predicts the relevance/irrelevance probabilities of an item, with uncertain items being considered serendipitous. Some interesting strategies have come from graph theory as well. Onuma et al. [24] proposed building a user-item graph with nodes receiving "bridging scores", where a high bridging score is assigned to nodes joining separate interconnected regions. Nakatsuji et al. [22] proposed a random walk on a user similarity graph to find another related user who might be a good source of serendipitous recommendations. On a different note, Zhang et al. [31] presented a system that uses Latent Dirichlet Allocation (LDA) to group Last.fm users into clusters, and simultaneously represent artists as a distribution over these clusters. This allows them to vectorize artists, and recommend items that are outside of a user's "musical bubbles". Lastly, we took inspiration from Adamopoulos and Tuzhilin [1], who proposed a method for generating serendipitous items based on how distant they are from a set of expected items. They formulate the utility function as a linear combination of an item's relevance score (coming from a baseline recommender) and the unexpectedness of the item (represented as the distance from a set of expected items). This distance between an item and set is calculated either by averaging all the individual distances or by calculating the distance from the centroid of the set. In our implementation, however, we take the minimum distance as detailed in Section 3.3.3.

2.3 Diversity-improving recommenders

A popular method of incorporating diversity is the greedy re-ranking approach. The goal here is to form a smaller subset of items (R) of size N from a larger set of candidates (C). It is an iterative algorithm and in each step, a candidate from C is chosen that maximizes the objective function, and is added to the result set R. This is repeated until R reaches the desired size N. The objective function is usually a linear combination between item relevance (predicted by a baseline recommender) and relative diversity with respect to items already in the set R. This relative diversity has been adapted by Kelly and Bridge [15], Smyth and McClave [26], and Ziegler et al. [33] as the average distance of an item from the items already in the result set R. Smyth and McClave [26] applied it to case-based recommenders, Ziegler et al.

[33] applied it to book recommendations, and Kelly and Bridge [15] used it in conversational recommender systems. This idea has been further developed by Jambor and Wang [13] to treat it as a constrained optimization problem, with relevance being the main objective and other beyond-accuracy metrics being additional constraints. Taking on a different perspective to the diversity problem, there exists another line of research that focuses on directly optimizing for diversity without using the greedy, "black-box" methods proposed above [11, 25, 27]. However, we use the greedy diversification algorithm in our study, which is described in more detail in Section 3.3.2.



Please compare the lists shown and answer the questions below:

1. Has a more personalized set of movies:

○ Definitely First ○ ○ Neither ○ ○ Definitely Second

2. Has more movies I haven't heard of:

○ Definitely First ○ ○ Neither ○ ○ Definitely Second

3. Better for picking a single movie to watch now:

○ Definitely First ○ ○ Neither ○ ○ Definitely Second

4. Better for picking a set of movies to watch over the next few weeks:

○ Definitely First ○ ○ Neither ○ ○ Definitely Second

5. Has a greater variety of movies:

○ Definitely First ○ ○ Neither ○ ○ Definitely Second

6. Better when looking for movies to watch with friends or family:

○ Definitely First ○ ○ Neither ○ ○ Definitely Second

Fig. 1. List comparison pages in the study look like this. Left pane has two recommendation lists: First and Second. Right pane has 11 questions. The image does not cover the full list of recommendations or questions here. Refer to Table 1 for the full list.

3 STUDY DESIGN

We conducted an online study for about 6 weeks (1.5 months) to gather user feedback on the recommendation lists presented to them and their general movie-watching preferences. The study was hosted on our movie recommendation website, MovieLens (https://movielens.org/)[1], which every user with >= 10 prior ratings had access to. This requirement for a minimum of 10 ratings was made in light of the serendipity algorithm needing prior ratings to work appropriately. The study spans 7 pages overall, and can be divided into two broad sections, as detailed below.

[1] MovieLens is a non-commercial, personalized recommender that gathers users' ratings on movies they have watched and provides predictions of new movies users might like to watch with different recommendation algorithms. The authors would like to thank the MovieLens team for their support and resources

	Question Type	*Purpose*	Question
L1	M	*Accuracy*	Has a more personalized set of movies
L2	M	*Novelty*	Has more movies I haven't heard of
L3	C	*Short-term goals*	Better for picking a single movie to watch now
L4	C	*Long-term goals*	Better for picking a set of movies to watch over the next few weeks
L5	M	*Diversity*	Has a greater variety of movies
L6	C	*Group watch*	Better when looking for movies to watch with friends or family
L7	P	*Dislike Impact*	Has more movies I dislike
L8	C	*Popularity*	Has more popular movies
L9	M	*Serendipity*	Has more movies I wouldn't have thought of myself, but think I might like
L10	P	*Preference level*	Overall, which recommendation list would you prefer
L11	C	*Context Feedback*	Are there situations you'd prefer a list different from the ones above? If so, when?

Table 1. List Comparison questions are grouped into (M)etrics, (C)context, and (P)reference based questions. L1-L10 ask users to compare the two lists shown, and give their preference on a 1-5 Likert scale.

3.1 List Comparison questions

It spans the first 6 pages in the study: each containing two recommendation lists (named First and Second) on the left pane, and 11 associated questions on the right as shown in Figure (1). For each page we pick a pair of algorithms from the Serendipity-improving (S), Diversity-improving (D), and Baseline (T) algorithms, (see Section 3.3) to generate our recommendation lists: so we formed 3 possible pairs (S&T, S&D, and D&T) and repeated them twice to use one for each of the 6 pages. By repetition we mean that the pairing combination was re-used and not the recommendations themselves: For example, Page 3 and Page 6 both see recommendations from the D&T pair, but we select a different, non-overlapping set of recommendations for each page. Furthermore, we also swap the order of the First and Second lists when a pair is repeated: For instance, in Page 3: First list is T and Second is D, while in Page 6: First is D and Second is T.

The 11 questions in turn asks the user's preference for the lists based on some criteria as shown in Table 1. Each of the first 10 questions presents them with a 5-point Likert scale spanning from 'Definitely First' to 'Definitely Second'. The final question (L11) has a text-response for user feedback on any additional context. The questions are designed to help us understand user perceptions of the algorithms on three fronts:

1: *Metrics (**M**)*: Accuracy (L1), Novelty (L2), Diversity (L5), Popularity (L8), and Serendipity (L9).

2: *Context (**C**)*: Short-term goals (L3), long-term goals (L4), and Group-watch (L6).

3: *Preferences (**P**)*: Dislikes (L7), Overall Preference (L10)

The questions for Serendipity (L9) and Diversity (L5) were inspired from our understanding of existing works [6, 20, 32]. The context-based questions were derived from pilot interviews for some of our earlier studies in the lab, during which users indicated a difference in expectations when watching with family or when picking movies to watch later, the latter commonly referred to as "stocking". Finally, with the dislike preference question (L7), we wanted to see if dislike for recommendations (or strong negative feedback) can have a stronger impact on preference compared to accuracy.

3.2 User Profile questions

It spans 1 page (the final page) and contains questions about the user's general movie-watching habits as shown in Table 2. The first question (U1) aims to understand the movie-watching frequency of the user. Note that this may

	Question	Options	Option Type
U1	How often do you watch movies? (Movies per month)	1-2 3-5 5-10 10-20 20+	Single Select
U2	How do you watch movies?	Exclusively in theatres Mostly in theatres Even mix of theatres and home viewing Mostly at home Exclusively at home	Single Select
U3	How would you describe your taste in movies?	Very mainstream Somewhat mainstream Somewhat eclectic Very eclectic	Single Select
U4	How would you describe your range of taste?	Very broad Somewhat broad Somewhat narrow Very narrow	Single Select
U5	Which of these options best match your goal on MovieLens?	Find movies I never heard of (*Novelty*) Find movies I heard of but forgot (*Unexpectedness*) Find movies similar to what I watched before (*Similarity*) Find a new variety of movies (*Diversity*)	Rate each from 1-5
U6	In your own words, please describe how you like to use MovieLens.	(Text Box Input)	Text Box

Table 2. User preference questions. Note: The goal purposes in U5's options are not shown to the users

not correspond to their log-in frequency on MovieLens. Through questions U3 & U4 we tried to understand how much they desired popular movies compared to niche ones, and see if they liked a broad range of content instead of narrow. Question U5 asks the user to rate each of the 4 possible goals they might have on MovieLens namely: novelty, unexpectedness, similarity, and diversity, on a scale of 1-5 based on how much they matter to them. We wanted to see if and how these user goals would impact user responses while comparing the recommendation lists. The last question is a text response, asking them for their overall feedback on their Movielens experience.

3.3 Algorithms used

Three different algorithms were used to generate recommendations in the study, and all of them were built on an ITEM-ITEM CF backbone. The algorithms have been tried and tested in prior research, and we chose them for their proposed effectiveness and ease of integration with our existing MovieLens platform. We also implemented a few software optimizations, like splitting the data load into multiple requests and utilizing memoization, to fetch the recommendations in a timely manner.

3.3.1 Baseline recommender (T). This corresponds to an existing algorithm in MovieLens that powers the 'Top-picks' page's recommendations. Although the recommendation engine for 'Top-picks' is customizable, allowing a user to switch between Baseline, SVD, and ITEM-ITEM CF engines, we stick with the ITEM-ITEM CF algorithm in our study. We use a popularity-based re-ranker to rank a bunch of candidates, as a linear combination between the ITEM-ITEM algorithm and a popularity-based ranker, that simply ranks movies based on popularity.

3.3.2 Diversity-improving recommender (D). This algorithm recommends a wider variety of movies than normal. Inspired from Kelly and Bridge [15], Smyth and McClave [26], Ziegler et al. [33], it is implemented as a diversity-based re-ranker, which is a linear combination between an ITEM-ITEM CF algorithm and a diversity ranker, as shown below:

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

where C indicates the list of candidate recommendations generated by the CF algorithm, and rel(i) is the prediction generated by it. This is to produce diverse recommendations while maintaining relevance. The latter represents movies as tag-genome vectors [18, 28] and measures each movie's average similarity with the rest of the candidates. Higher similarity gets a lower score, and the scores are normalized before combining linearly. The main idea here is to give higher priority to movies that are different from the rest of the candidates. Cosine similarity is used as the distance metric in the score calculation.

3.3.3 Serendipity-improving recommender (S). This algorithm is a serendipity-improving recommender. Inspired from Adamopoulos and Tuzhilin [1], it is implemented as a serendipity-based re-ranker, which is a linear combination between an ITEM-ITEM CF algorithm and a serendipity ranker. The latter ranks candidates by representing them as tag-genome vectors and measuring each candidate's distance to its closest neighbor among the user's previously rated movies, as shown below:

$$f_{ser}(i, R) = w_s.\text{rel}(i) + (1 - w_s).\min_{j \in C} \text{dist}(i, j) \tag{eq.2}$$

where C indicates the list of most recently rated 200 movies, and rel(i) is the prediction generated by the CF algorithm. Higher distance gets a higher score, and the scores are normalized before combining linearly. The main idea is to find movies that are farther from the user's current tastes while trying to maintain relevance. The distance to the closest neighbor was chosen in accordance to [14], which indicated that choosing the minimum distance could be more appropriate than averaging it, in order to avoid losing information, especially with diverse users. As with the diversity-based ranker, Cosine similarity is used as a distance metric.

3.3.4 Pilot study and algorithm validation. To validate the algorithms implemented above, and choose the appropriate weights (w_d, w_s) for the linear combination in each of the re-rankers, we conducted a pilot study with 4 users. The pilot users were shown two sets of recommendations with 7 recommendation lists in each, where each list contains 16 movies. One set of recommendations varied from low serendipity to high, while the other set varied from low diversity to high, generated by adjusting the algorithm weights. The lists were presented to the users in a random order, and they were asked to arrange the corresponding lists in increasing order of serendipity and diversity. Averaging the result from all the users we found that w_s=0.10 (0.90 * Baseline score + 0.10 * Serendipity score) worked best for the Serendipity ranker, and w_d=0.30 (0.70 * Baseline score + 0.30 * Diversity score) worked best for the Diversity ranker. For the Serendipity ranker, only a small w_s-step away from the baseline was considered serendipitous. For Diversity, user perception of variety seemed to increase up to a few w_d-steps and then stabilized, as shown by Willemsen et al. [29] as well. We also experimented with different distance formulas, by manually evaluating the recommendations produced from them. We observed that using dot and eucledian products produced too similar or unpopular movies. Jaccard similarity worked well with genre-based vectors instead of tag-genomes. Scaled Dot product gave better results but introduced a new parameter (scale) to adjust. Cosine similarity seemed to be the best for both Diversity and Serendipity.

With Diversity it's able to recommend shows and stand-ups in addition to movies, and the predicted ratings of these movies are also higher, indicating that it's able to produce more relevant movies.

4 RESULTS

4.1 List comparison overview (Identifiability)

We evaluated the responses of L9 (Serendipity) and L5 (Diversity) for the three recommender pairs to understand how the users perceived them. As seen in Table (3), users' perception of Diversity matches with our expectations when paired with either S or T recommenders i.e they find D to be more diverse. Comparing the distribution of L5 responses for D&T and D&S pairs, we found no significant difference (p-value = 0.82676). This indicates that D is perceived as more diverse in both the cases, although it is only slightly more diverse than T. The scenario is a little more interesting for Serendipity, where users seem to find T more serendipitous when S & T are paired, and S more serendipitous when S & D are paired. Comparing the distribution for these two pairs does show a significant difference (p-value = 0.05724). A more detailed exploration of this can be found in Section 4.4.2.

Diversity	D > T	D < T	D = T	D > S	D < S	D = S
	413	405	411	384	350	467
Serendipity	S > T	S < T	S = T	S > D	S < D	S = D
	325	346	802	361	316	273

Table 3. List comparisons: The first row compares the algorithms by counting user responses for diversity (L5), while the second row does it for serendipity (L9).

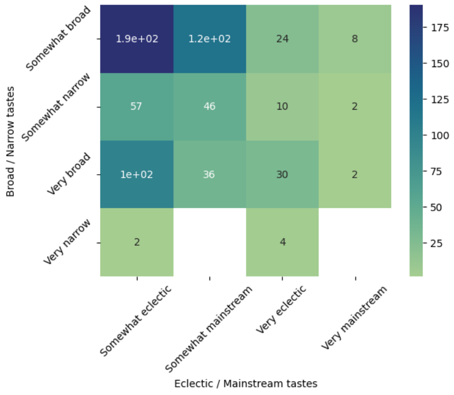


Fig. 2. Comparing user tastes

4.2 User profile overview

Many of the users seem to watch between 3-10 movies (U1). Around 90% of them reported saying they watch mostly at home or exclusively at home, the former being large (U2). Regarding tastes, a majority identified their tastes to be 'somewhat eclectic', followed by 'somewhat mainstream', with these two categories making up 87% of the users (U3). It's interesting to note that on the extreme ends, it seems like users tend to be 'very eclectic' (10%) much more often than 'very mainstream' (2%). On the breadth of their tastes, users majorly identified as 'somewhat broad' or 'very broad', together making up 81% of the users (U4). Only 1% of users identified their tastes to be 'Very narrow'. Analyzing Q3 and Q4 together (see figure (2)), we found that the most popular combination is 'somewhat broad' + 'somewhat eclectic', followed by 'very broad' + 'somewhat eclectic' and 'very broad' + 'somewhat mainstream'. Even among very eclectic users, 80% (36 / 46) prefer somewhat broad or very broad recommendations. This necessitates variety even among very eclectic tastes and pigeonholed users. On user goals: Finding movies they've never heard of (Novelty), and movies that are similar to what they've watched before (Similarity) take higher precedence. Finding movies they heard of but forgot (Unexpectedness), and a new variety of movies (Diversity) take lower precedence. It's interesting to note that users desire variety or broadness in their recommendations, but not necessarily a new variety of movies.

4.3 Text feedback overview

4.3.1 On MovieLens usages. U6 on the final user profiles page brought to light a variety of ways in which users liked to use MovieLens. A lot of the users use it for book-keeping purposes, "to keep a diary" of the movies they've watched. It was interesting to see some users engage in trust-building exercises, where they experimented with the algorithm by seeing how closely its predicted ratings matched with their own or those of the people they knew. This can be quite important as some users mentioned that they don't use MovieLens for recommendations, but for the predicted ratings to decide if they'd like the movie. Some users exhibit niche goals: they like to find lesser-known, non-mainstream, or hard-to-find movies and look for a specific genre or mood, or "dive into smaller genres". We also saw users describing exploration goals: "Find familiar movies, but branch out a bit", unexpectedness goals: "Rediscover movies that I haven't watched in a while and don't come to mind readily", and serendipity goals: "discover great films I would've otherwise never discovered on my own". Users describe the need for diversity in the context of group recommendations. These users use their accounts to get recommendations for themselves and their families, so they felt the need for a "broader recommendation algorithm that can recommend on several different subsets and genres rather than trying to match it all up as if it's for just one person". Since user goals are so different and vary based on context, it would be difficult to have a single algorithm optimized for them all. But providing them with better control can help better fulfill their goals instead. Users did indicate that the algorithm isn't perfect by itself, but all the search tools together help them reach their goal: "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". These controls can help users communicate their interests better, as highlighted by the discussion below.

4.3.2 On Negative Feedback. We made a special note on the need users felt for negative feedback since other forms of explicit feedback apparently weren't enough. For instance, a user mentioned that their low rating for a movie didn't necessarily mean they didn't like it, but that they didn't want to watch it now. A common request among a multitude of users is the ability to filter out movies they don't like. Some of this capability is straightforward: filtering out based on genre, time period, tags, or cast, while some is more nuanced: "already familar with USSR films, no need of suggestions. Won't watch any victorian or Indian film". It is important to note that MovieLens does provide the option to filter based on genre, time period, or tags, but these are to include the selected options not exclude them. One user requested the ability to refresh their recommendations. This ability, we feel, if implemented is best left to the user since automatic refresh can confuse them. In fact, one user mentioned they were "afraid" they were "going to miss movies because the recommendations update so frequently, and (they're) afraid it will never repeat a recommendation so (they) may have lost some entirely".

4.3.3 On List comparisons. Some users perceived the lists as we would expect them to. A user remarked, "The first (T) seems better for when I'm looking for a quick recommendation to start watching right away; the second (D) for if I'm looking to broaden my movie horizons and get recommendations I might not have considered". One user liked T because it had fewer movies they disliked. Another user when comparing T with D, felt T "followed (their) tastes very well", while D had "challenging/new content". The same user, on comparing D with S, felt D had more movies they liked, while S had "a mix of things (they) like and would probably like", defining Serendipity perfectly. Although, not all users had the same experience as many reported that they equally liked or disliked the lists presented. One user mentioned that the Serendipity list was "full of cheesy junk that I wouldn't watch".

On another note, we were glad to see some contexts emerge from the user responses, on when these lists could be helpful to them or when a different list might be useful. A user mentioned T was "great for cozy watching when you want something that is definitely good", while D was "good for long-term planning...or when you want to be challenged by something you've never heard of right now...with a bigger focus on accessible family content". The user wished for a mix of both of those algorithms and not just one. Some users mentioned they would prefer one list over the other "when looking for lighter movies", "when looking for more modern movies", or for "discovery of lesser-known movies", although they didn't mention which list they were referring to.

When asked to compare the lists, it was interesting to see users talk about movie diversity by considering genres, release years, and sometimes tags. Different users saw it differently, suggesting Diversity is user-specific and a "unified" definition for it may be hard to derive.

	Question purpose	Ranking
L1	Accuracy	T >>>> D >>>>> S
L2	Novelty	S >>>>>>>>> D > T
L3	Short-term goals	T >>>> D >>>>>> S
L4	Long-term goals	T >>> D >>>>> S
L5	Diversity	T > D >>> S
L6	Group watch	T > D >>>>>> S
L7	Dislike	S > D >>> T
L8	Popularity	T > D >>>>>>>>>>> S
L9	Serendipity	T >> S > D
L10	Preference	T >>>> D >>>> S

Table 4. Count Comparison: The number of > are proportional to the gap between recommender votes (~30 for each >)

4.4 Analysis deep-dive

4.4.1 Count comparisons. We carried out a count-based comparison between the three recommenders for all the metric (M), context (C), and preferences (P) questions to see how the algorithms rank. T remains the most preferred and accurate recommender overall, followed by D and S in that order with sizeable gaps between each. To our mild surprise, T was also perceived to be much more serendipitous, with S & D being close. T & D are close in perceived diversity, which is much more than that for S. On the other hand, S & D are much more disliked than T. S is also perceived as the most novel and least popular, with T & D being close. In essence, S is the most novel, yet most disliked, least accurate, and least popular. So it seems to have traded-off novelty for relevance, as expected, and is disadvantaged in terms of the "sense of familiarity" or trust generally induced by familiar movies [2], making it harder for users to decide if they would like them. A user could be quoted as saying it is "really hard to compare as both lists are mostly films I don't know much about", which could be especially true for novel items. This could be one of the reasons why T outperformed S in serendipity. Taking inspiration from a user who suggested a list by "moving slightly away from their wishlist along a particular genre or actor", we believe taking smaller steps along preferred directions could be a good way to let users discover possibly serendipitous items, as already shown by Nakatsuji et al. [22]. Interestingly, when only T & D are compared, they perform similarly on all regards.

T, D, and S are preferred for long-term and short-term goals in that order, with sizeable gaps in between. It's interesting to note that the gap between T and D is smaller for long-term goals, indicating that users might want more diverse options when making a list to watch later. D is very close to T for group watch, while S lags miles behind. Users

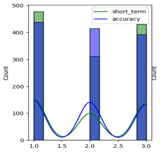
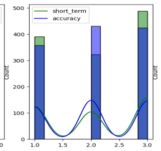
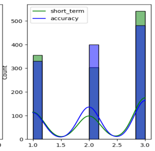
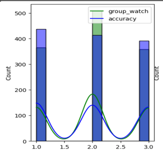
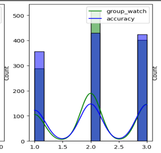
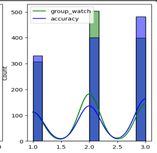
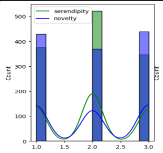
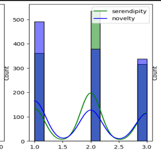
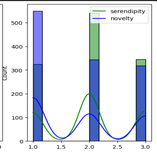
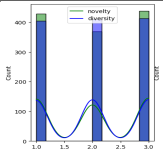
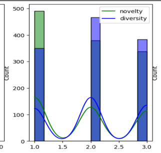
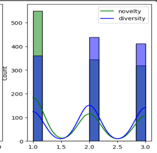
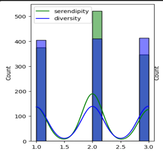
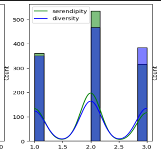
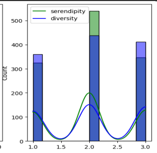
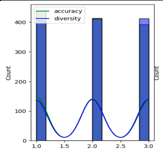
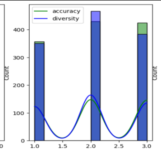
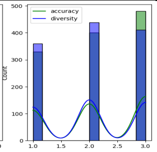
Metrics compared	T & D	S & D	S & T
Short-term vs Accuracy	 0.580	 0.532	 0.495
Group-watch vs Accuracy	 0.150	 0.258	 0.160*
Novelty vs Serendipity	 0.089	 0.103**	 0.051**
Novelty vs Diversity	 0.021	 0.009**	 0.022**
Serendipity vs Diversity	 0.170	 0.159**	 0.158
Accuracy vs Diversity	 0.135	 0.153	 0.110**

Table 5. Metrics Comparison: The plots show user response distributions for each recommender pair. The responses 1-2 (First preferred) have been grouped as 1, 3 (Neutral) as 2, and 4-5 (Second preferred) as 3. The values below the figure indicate Spearman's rank order correlation values (ρ). A * indicates p-value < 0.1, and ** indicates p-value < 0.05 from a Mann-Whitney U test for a difference between the distributions.

seem to prefer the diverse recommender when watching with groups or family, as explicitly mentioned by some users (Refer to Sections 4.3.3, 4.3.1). In absolute terms, T checks many of the boxes whereas S and D seem more use-case specific so they could prove useful despite being less preferred or accurate.

4.4.2 Metric Interactions. We conducted a few tests to see how these metrics (M), contexts (C), and preferences (P) interact with and influence each other. Two of them are compared at a time by taking user responses for each recommender pair individually, and testing for similarity using a Mann-Whitney U test. A handful of these comparisons are shown in Table 5.

Short-term: It doesn't seem to be associated with serendipity (L9), while it is negatively with novelty (L2). It is positively associated with perceived accuracy (L1), overall preference (L10), popularity (L8), and weakly positively with diversity (L5). Getting it right seems important for picking a movie to watch now.

Long-Term: It is negatively associated with novelty. Positively with accuracy, overall preference, and weakly positively with diversity, serendipity, and popularity. Interestingly, we found that within each recommender pair, the difference in votes between the recommenders is larger for accuracy (L1) and popularity (L8) questions than that for long-term usage (L4) when comparing L1 vs L4, and L8 vs L4. So the gap seems to narrow for long-term usage, hinting at the trade-offs users seem to make for long-term movie picking.

Group-watch: It is weakly positively associated with diversity, popularity, overall preference, and accuracy. Like the others, it is negatively associated with novelty. Although, it seems to be unaffected by serendipity. Interestingly, its association with perceived accuracy and preference is very low compared to that for long-term and short-term goals, as indicated by Spearman's rho (ρ) values in Table 5. So for watching with friends and family personalization could be compromised with diversity, and serendipity may not be a good choice here as getting it right could be a little tricky.

Algorithms: Comparing responses for S & D: S is seen as more novel and serendipitous, while D is seen as more diverse, popular, accurate, and preferred overall. D & T are seen as equal performers in all those regards. Comparing responses for S & T: T beats S in all aspects, except for novelty which is very high for S. (see Novelty vs Serendipity, Serendipity vs Diversity in Table 5)

When D & T perform almost similarly, why do we see a difference when comparing them with S? Well, D is slightly more diverse of the two, while T is slightly more accurate (see Accuracy vs Diversity in Table 5). We did find a significant difference between the ILS scores of the recommendations shown by D & T (p-value=0.090, alpha=0.1), with the mean ILS of T being higher. As T is less diverse and more personalized, it might be accentuating any differences in novelty. We did indeed find that S is perceived as way more novel when compared with T than with D. As T seems to have more movies that users like, it could be perceived as more serendipitous as well (More on this in Section 4.4.3).

Since we observed that novelty is negatively associated with many of the goals and preferences, we were motivated to check if novelty is particularly disliked by users. **Novelty vs Dislike**: We see that D & T are equally novel and equally disliked. For D & S, S is more novel, but they're similarly disliked. For S & T, S is more novel, and T has slightly fewer dislikes. Novelty, in general, is very weakly associated with dislike. As the goals clearly indicate (Section 4.2), novelty-seeking is an important goal, but they also want them to be relevant.

Knijnenburg et al. [16] conducted a user study to analyze diversification algorithms and found that user satisfaction is positively associated with diversity, and negatively with novelty. So we were curious to see if the two metrics interact in any way, specifically if novelty would make it difficult to perceive diversity. **Novelty vs Diversity**: As mentioned before - D & T are perceived as equally diverse and novel, For S & D: S is perceived as more novel, and D as more diverse. For S & T: S is perceived as way more novel, and T as more diverse. So it seems users can perceive novelty and diversity separately and, as we've seen, prefer them in different contexts. (see Novelty vs Diversity in Table 5)

Lastly, we analyze how popularity is associated with personalization and dislike. We found it is positively associated with accuracy and overall preference. And it is inversely related to novelty. But it is not associated with dislikes: so users do not necessarily hate recommendations that have popular movies. Overall, they want the movies to be somewhat mainstream (Section 4.2), so some popularity seems desirable.

4.4.3 User-type comparisons. In Section 4.4.2, we proposed that T is seen as more serendipitous because it's less diverse and more accurate than D. So we wanted to see if Diverse and Non-diverse users would perceive them differently.

For this, we calculated an ILS score for each user on the movies rated by them and split them into two groups at the median ILS score. The movies were rated by users on MovieLens before, and we represented them as tag-genome vectors. Once we had the groups, we were able to compare and contrast their perceptions of diversity, serendipity, accuracy, and novelty. From this, we found that the two groups perceived all the metrics similarly, except for serendipity. When comparing S & T, Diverse users see them as similarly serendipitous, with S being slightly higher, while Non-diverse users see T as being more serendipitous. Initially, we suspected that this could be due to diverse users having more variety in T, so novelty in S may not come across as badly for them in comparison. But we found no significant difference between the ILS scores of movies shown by T to diverse and non-diverse users. So it seems like a simple popularity mix, as in T, feels more serendipitous to non-diverse users, while for diverse users it is the serendipity-improving algorithm. This could be due to a natural inclination of diverse users to consume content with more variety.

We also wanted to check if frequent movie watchers have goals and perceptions that are different from in-frequent ones. For users who watch <=5 movies per month (U1), 'Somewhat Broad' and 'Somewhat Narrow' (U4) were the top two preferences. For >=10 movie watchers, 'Very Broad' takes the second position, and for 20+ movie watchers it takes the first. Since we did notice a difference in goals, we created two new groups: with <=10 movie watchers being called Infrequent, and the rest as Frequent. Between these groups, diversity goals (U5) seemed to be different (p-value: 0.0007), with Frequent users having a higher priority for broadness/variety, while the other goals of similarity, novelty, and unexpectedness (U5) seemed to be the same for both.

When comparing the responses of these groups for different metrics, we observed that the pairs S & T, and S & D were perceived similarly, but for D & T some perceptions were different. Frequent users saw D as more serendipitous, popular, and accurate (metrics that usually don't go together), while infrequent users thought it was T. Frequent users also saw D as more diverse and novel, although we didn't find a significant statistical difference from T. Nevertheless, both user groups prefer T overall, while D came in as a close second for frequent users. So it seems like the perceptions of Frequent and Infrequent users differ a lot, and frequent users could benefit from having a diversity-oriented recommender.

Thinking along the lines of broadness/variety, we also compared users for whom finding similar movies is a high priority (U5) with users for whom it is not. We're calling them similarity-seeking users, and non-similarity-seeking users. When comparing the responses for D vs S lists, we found that similarity-seeking users perceive D to be more accurate and preferred, while the other group thinks it's S, although both the groups perceived S to be more novel and serendipitous. It's understandable to see similarity-seeking users not like the level of novelty presented by S.

Furthermore, we divided the users into different groups based on their responses to the other user profile questions (see Table 2), and historical data: novelty-seeking vs non-novelty-seeking users (U5), eclectic vs mainstream users (U3), fewer ratings vs more ratings users (Based on their historical rating data in MovieLens), and expected goal vs unexpected goal users (U5). On comparing their perceptions of different metrics, we found no differences for any given pair. But we did observe that the perceptions sometimes differed when the S recommender was involved, although they were similar for most of the other metrics. Based on this, we have reason to believe that the type of user to which "serendipitous" recommendations are shown matters.

5 CONCLUSION

We observed a lot of differences in perception, some of them being very interesting. Users seem to perceive novelty, popularity, and diversity clearly, but diversity can be much more nuanced and tougher to generalize. For instance, a movie that is rated by a large number of users can be safely considered popular. But a variety of fantasy/fiction movies may not be perceived as diverse since they're all from the same genre. Some users perceive diversity based on genres, some on release years, some on cast, tags, etc, so it needs to be personalized. In fact, we find that even users with niche tastes look for variety. So they probably don't intend to broaden their tastes but want variety within its confines to help them find something they might like.

Diversity also seems to play a role when people are picking something to watch with their friends and family. In such cases, the expectation for personalization is lower, and it is higher for variety. The algorithm is usually unaware of this context so it tries to tie all these tastes together without realizing that the user actually sees them as separate objectives.

Another shortcoming we identified in MovieLens, and recommender systems in general, is an effective incorporation of negative feedback. We found that users felt the urge to "tell" the recommender what they do not want. A majority wanted to filter out movies based on genres, actors, or release years. Current MovieLens controls allow users to filter their search by *including* these attributes, but not *excluding* them. Without appropriate controls, users are forced to adapt in ways that are confusing for both the developers and the algorithm. For instance, some users gave low ratings when they didn't necessarily hate a movie but just wanted to watch it later. A few users tend not to rate too many movies in a certain genre to avoid flooding their recommendations with them. This calls for better controls that can put users in charge so they can get the content they want.

Finally, we note that the serendipity-improving algorithm (S) failed to produce the desired effect, and a major reason for this, we believe, is the way it is formulated. Since it looks at the closest neighbor's distance while ranking an item, it can be prone to ignore the overall watch frequencies of a user. Consider the scenario where a movie (say "A") is very close to only one of the user's previously rated movies (say "B"), and is quite far from the rest. This could be a relevant and potentially surprising recommendation to the user, but since it is very similar to "B", our algorithm won't prioritize it. The original implementation by Adamopoulos and Tuzhilin [1] uses average distance instead of a minimum, and that might be better suited in such scenarios.

6 FUTURE IMPLICATIONS

In light of the differences in perception among different user groups, there's a need for a way to paint a better, more accurate picture of the user. For this, any additional information to fill the existing gaps could be useful, eg:- How many movies do they watch per month? Are they looking for movies similar to what they've seen, or something different? Are they looking for something to watch now or for later? But practically, it's not easy to gather this information. Even if we did, how often would we update it? Of all the logins at MovieLens, only 4-5% of them have seen similar movie search, and tag actions, while about 55-58% of them have had a rating, or a movie-page view related action. So users don't seem to want to do much beyond rate movies and watch descriptions.

In fact, some users were found complaining or requesting features that already exist on our website. When one of our pilot users wished for the ability to remove a recommendation from their list, we asked if they were aware of the "hide" option in our system that serves this purpose, to which they said they weren't. Maybe the user missed it because it was concealed behind a menu button. Another user said they were aware of this feature but didn't use it because they didn't fully understand how it worked: Does it hide only this movie? Or movies *like* this? Unintuitive design can hurt user experience. We need design that is simple, intuitive, and clear to the user, so that it is useful and transparent.

A lot of users felt the "Similar movies" section in MovieLens had good recommendations. In fact, some use it primarily for exploration. This closely relates to another user's request for a list that deviates only slightly from their wishlist along a particular genre. The algorithm (S) used in our analysis tries to move furthest from the user's current tastes while balancing relevance, causing some recommendations to appear "too wild". It was able to generate perceivably novel items but failed to be relevant enough to be serendipitous. The experiment accentuates the challenges with serendipity and suggests smaller "steps" from a user's current interests as a possible alternative.

More so, every application might need separate strategies and controls as different types of users perceive things differently. For instance, users on a non-profit, research-oriented movie recommender system, like MovieLens, might have a higher tolerance for error compared to an average user on a commercial movie recommender, like Netflix[2]. Users on a professional networking platform, like LinkedIn[3], might be more tech-savvy than the average individual so a more nuanced recommender control might work fine here. It would be interesting to see how different user groups differ in their needs, and how we could develop strategies to serve them accordingly.

REFERENCES

[1] Panagiotis Adamopoulos and Alexander Tuzhilin. 2014. On Unexpectedness in Recommender Systems: Or How to Better Expect the Unexpected. *ACM Trans. Intell. Syst. Technol.* 5, 4, Article 54 (dec 2014), 32 pages. https://doi.org/10.1145/2559952

[2] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. 2017. How to Recommend? User Trust Factors in Movie Recommender Systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 287–300. https://doi.org/10.1145/3025171.3025209

[3] Òscar Celma. 2010. *Music Recommendation and Discovery*. Springer-Verlag, Berlin, Heidelberg. 194 pages. https://doi.org/10.1007/978-3-642-13287-2

[4] Li Chen, Yonghua Yang, Ningxia Wang, Keping Yang, and Quan Yuan. 2019. How Serendipity Improves User Satisfaction with Recommendations? A Large-Scale User Evaluation. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 240–250. https://doi.org/10.1145/3308558.3313469

[5] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide amp; Deep Learning for Recommender Systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems* (Boston, MA, USA) *(DLRS 2016)*. Association for Computing Machinery, New York, NY, USA, 7–10. https://doi.org/10.1145/2988450.2988454

[6] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2014. User Perception of Differences in Recommender Algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems* (Foster City, Silicon Valley, California, USA) *(RecSys '14)*. Association for Computing Machinery, New York, NY, USA, 161–168. https://doi.org/10.1145/2645710.2645737

[7] Florent Garcin, Boi Faltings, Olivier Donatsch, Ayar Alazzawi, Christophe Bruttin, and Amr Huber. 2014. Offline and Online Evaluation of News Recommender Systems at Swissinfo.Ch. In *Proceedings of the 8th ACM Conference on Recommender Systems* (Foster City, Silicon Valley, California, USA) *(RecSys '14)*. Association for Computing Machinery, New York, NY, USA, 169–176. https://doi.org/10.1145/2645710.2645745

[8] Mouzhi Ge, Fatih Gedikli, and Dietmar Jannach. 2011. Placing high-diversity items in top-N recommendation lists. *Proceedings of the 9th Workshop on Intelligent Techniques for Web Personalization Recommender Systems at IJCAI 2011* (01 2011), 65–68.

[9] Mouzhi Ge, Dietmar Jannach, and Fatih Gedikli. 2013. Bringing Diversity to Recommendation Lists – An Analysis of the Placement of Diverse Items. In *Enterprise Information Systems*, José Cordeiro, Leszek A. Maciaszek, and Joaquim Filipe (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 293–305.

[10] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Trans. Inf. Syst.* 22, 1 (jan 2004), 5–53. https://doi.org/10.1145/963770.963772

[11] Neil J. Hurley. 2013. Personalised Ranking with Diversity. In *Proceedings of the 7th ACM Conference on Recommender Systems* (Hong Kong, China) *(RecSys '13)*. Association for Computing Machinery, New York, NY, USA, 379–382. https://doi.org/10.1145/2507157.2507226

[12] Leo Iaquinta, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, Michele Filannino, and Piero Molino. 2008. Introducing Serendipity in a Content-Based Recommender System. In *2008 Eighth International Conference on Hybrid Intelligent Systems*. 168–173. https://doi.org/10.1109/HIS.2008.25

[2] https://www.netflix.com/

[3] https://www.linkedin.com/

[13] Tamas Jambor and Jun Wang. 2010. Optimizing Multiple Objectives in Collaborative Filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (Barcelona, Spain) *(RecSys '10)*. Association for Computing Machinery, New York, NY, USA, 55–62. https://doi.org/10.1145/1864708.1864723

[14] Marius Kaminskas and Derek Bridge. 2016. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Trans. Interact. Intell. Syst.* 7, 1, Article 2 (dec 2016), 42 pages. https://doi.org/10.1145/2926720

[15] John Kelly and Derek Bridge. 2006. Enhancing the diversity of conversational collaborative recommendations: A comparison. *Artif. Intell. Rev.* 25 (04 2006), 79–95. https://doi.org/10.1007/s10462-007-9023-8

[16] Bart Knijnenburg, Martijn Willemsen, gantner, soncu, and newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22 (10 2012), 441–504. https://doi.org/10.1007/s11257-011-9118-4

[17] Denis Kotkov, Joseph A. Konstan, Qian Zhao, and Jari Veijalainen. 2018. Investigating Serendipity in Recommender Systems Based on Real User Feedback. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (Pau, France) *(SAC '18)*. Association for Computing Machinery, New York, NY, USA, 1341–1350. https://doi.org/10.1145/3167132.3167276

[18] Denis Kotkov, Alexandr Maslov, and Mats Neovius. 2021. Revisiting the Tag Relevance Prediction Problem. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1768–1772. https://doi.org/10.1145/3404835.3463019

[19] Denis Kotkov, Jari Veijalainen, and Shuaiqiang Wang. 2016. Challenges of Serendipity in Recommender Systems. 251–256. https://doi.org/10.5220/0005879802510256

[20] Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. 2016. A survey of serendipity in recommender systems. *Knowledge-Based Systems* 111 (2016), 180–192. https://doi.org/10.1016/j.knosys.2016.08.014

[21] Tomoko Murakami, Koichiro Mori, and Ryohei Orihara. 2007. Metrics for Evaluating the Serendipity of Recommendation Lists. In *Proceedings of the 2007 Conference on New Frontiers in Artificial Intelligence* (Miyazaki, Japan) *(JSAI'07)*. Springer-Verlag, Berlin, Heidelberg, 40–46.

[22] Makoto Nakatsuji, Yasuhiro Fujiwara, Akimichi Tanaka, Toshio Uchiyama, Ko Fujimura, and Toru Ishida. 2010. Classical Music for Rock Fans? Novel Recommendations for Expanding User Interests *(CIKM '10)*. Association for Computing Machinery, New York, NY, USA, 949–958. https://doi.org/10.1145/1871437.1871558

[23] Jinoh Oh, Sun Park, Hwanjo Yu, Min Song, and Seung-Taek Park. 2011. Novel Recommendation Based on Personal Popularity Tendency. In *2011 IEEE 11th International Conference on Data Mining*. 507–516. https://doi.org/10.1109/ICDM.2011.110

[24] Kensuke Onuma, Hanghang Tong, and Christos Faloutsos. 2009. TANGENT: A Novel, 'Surprise Me', Recommendation Algorithm. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Paris, France) *(KDD '09)*. Association for Computing Machinery, New York, NY, USA, 657–666. https://doi.org/10.1145/1557019.1557093

[25] Yue Shi, Xiaoxue Zhao, Jun Wang, Martha Larson, and Alan Hanjalic. 2012. Adaptive Diversification of Recommendation Results via Latent Factor Portfolio. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Portland, Oregon, USA) *(SIGIR '12)*. Association for Computing Machinery, New York, NY, USA, 175–184. https://doi.org/10.1145/2348283.2348310

[26] Barry Smyth and Paul McClave. 2001. Similarity vs. Diversity. In *Proceedings of the 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development (ICCBR '01)*. Springer-Verlag, Berlin, Heidelberg, 347–361.

[27] Ruilong Su, Li'Ang Yin, Kailong Chen, and Yong Yu. 2013. Set-Oriented Personalized Ranking for Diversified Top-n Recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems* (Hong Kong, China) *(RecSys '13)*. Association for Computing Machinery, New York, NY, USA, 415–418. https://doi.org/10.1145/2507157.2507207

[28] Jesse Vig, Shilad Sen, and John Riedl. 2012. The Tag Genome: Encoding Community Knowledge to Support Novel Interaction. *ACM Trans. Interact. Intell. Syst.* 2, 3, Article 13 (sep 2012), 44 pages. https://doi.org/10.1145/2362394.2362395

[29] Martijn Willemsen, Bart Knijnenburg, Mark Graus, Linda Velter-Bremmers, and Kai Fu. 2011. Using latent features diversification to reduce choice difficulty in recommendation lists. *CEUR Workshop Proceedings* 811.

[30] Liang Zhang. 2013. The Definition of Novelty in Recommendation System. *Journal of Engineering Science and Technology Review* 6 (06 2013), 141–145. https://doi.org/10.25103/jestr.063.25

[31] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. 2012. Auralist: Introducing Serendipity into Music Recommendation *(WSDM '12)*. Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/2124295.2124300

[32] Qian Zhao, Gediminas Adomavicius, F. Maxwell Harper, Martijn Willemsen, and Joseph A. Konstan. 2017. Toward Better Interactions in Recommender Systems: Cycling and Serpentining Approaches for Top-N Item Lists. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1444–1453. https://doi.org/10.1145/2998181.2998211

[33] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving Recommendation Lists through Topic Diversification *(WWW '05)*. Association for Computing Machinery, New York, NY, USA, 22–32. https://doi.org/10.1145/1060745.1060754