MovieTweeters

An Interactive Interface to Improve Recommendation Novelty

Ghanmode, Ishan ; Tintarev, Nava

Publication date
2018

Document Version
Accepted author manuscript

Published in
Proceedings of the 5th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems

Citation (APA)
Ghanmode, I., & Tintarev, N. (2018). MovieTweeters: An Interactive Interface to Improve Recommendation Novelty. In P. Brusilovsky, M. de Gemmis, A. Felfernig, P. Lops, J. O'Donovan, G. Semeraro, & M. C. Willemsen (Eds.), *Proceedings of the 5th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems* (pp. 24-31). (CEUR Workshop Proceedings; Vol. 2225). CEUR-WS. http://ceur-ws.org/Vol-2225/

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

MovieTweeters

An Interactive Interface to Improve Recommendation Novelty

Ishan Ghanmode
Delft University of Technology
Delft, The Netherlands
I.Ghanmode@student.tudelft.nl

Nava Tintarev
Delft University of Technology
Delft, The Netherlands
n.tintarev@tudelft.nl

ABSTRACT

This paper introduces and evaluates a novel interface, *MovieTweeters*. It is a movie recommendation system which incorporates social information with a traditional recommendation algorithm to generate recommendations for users. Few previous studies have investigated the influence of using social information in interactive interfaces to improve the novelty of recommendations. To address this gap, we investigate whether social information can be incorporated effectively into an interactive interface to improve recommendation novelty and user satisfaction. Our initial results suggest that such an interactive interface does indeed help users discover more *novel* items. Also, we observed users who *perceived* that they discovered more *novel* and *diverse* items reported increased levels of *user satisfaction*. Surprisingly, we observed that even though we successfully were able to increase the *system diversity* of the recommendations, it had a negative correlation with users perception of *novelty* and *diversity* of the items highlighting the importance of improved user-centered approaches.

CCS CONCEPTS

• **Human-centered computing** → **User interface design**; • **Information systems** → **Recommender systems**;

KEYWORDS

Social Recommendation Systems, Novelty, Diversity, User Satisfaction, Interactive User Interfaces

1 INTRODUCTION

Social networks such as Facebook[1] and Twitter[2] have emerged as some of the most popular social media platforms allowing users to communicate and express their opinions and feedback. This online interaction between users generates social and preference information which could be effectively harnessed in recommender systems. This social information has been found to improve algorithmic performance [28]. However, accurate recommendations do not always correspond to higher levels of user satisfaction [22, 27]. In response, researchers have proposed *'beyond accuracy'* metrics such as *diversity* and *novelty* [10, 26], and worked on interfaces to improve the quality of recommendations [27].

A limited, but growing, body of literature has studied the influence of using *social information* in interactive interfaces to improve the novelty of recommendations. To address this gap, we investigate whether social information can be incorporated effectively into an interactive interface to improve recommendation novelty. Our contributions are as follows:

- We introduce a novel interface, *MovieTweeters*, which incorporates social information into a traditional recommendation algorithm. This enables users to leverage their relevant social information and discover novel (and more recent) content.
- This paper evaluates the system in terms of its ability to improve: a) *system diversity*; b) *perceived novelty*, c) *perceived diversity*.
- We study the relationship between system and user measures. We also establish a positive impact of users perceived quality of recommendations on their overall satisfaction.

The remainder of this paper is organized as follows: First, we present a discussion of related work in Section 2. Next, we introduce the *MovieTweeters* system, including the design choices for the interactive user interface in Section 3. We also discuss the underlying algorithms used in the system. In Section 4, we describe an online user experiment (N=23) in which we evaluate the system. We present our results in Section 5. A brief discussion of the notable results are presented in Section 6 and finally we conclude in Section 7 with ideas for future work.

2 RELATED WORK

To frame our research done in this paper in terms of related work, we discuss three key areas in detail. First, we discuss related work in existing *social recommendation systems*. Second, we focus on the importance of *inspectability* and *control* on recommendation system interfaces. We also look into how these interfaces have an impact on the users. Finally, we also present a discussion of related work in the area of *beyond accuracy metrics* such as *diversity* and *novelty*.

2.1 Social Recommendation Systems

One of the definitions for social recommendation is any given recommendation with online social information as an additional input i.e augmenting or improving the existing social recommendations with additional social information [13]. One of the earliest works which included social properties was done in [12], where the researchers built *ReferralWeb*. It was an interactive system for searching relevant social networks on the World Wide Web. Social information can be in the form of social relations, friendships, social influence and so on. In this definition, the social recommendation systems assume that the users are related when they establish social relations. Under this assumption, the social information or social relations are used to improve the performance of the recommendations [19]. We base our study around the first definition, where we use existing social information as an additional input to improve the quality of recommendations.

[1] https://www.facebook.com, accessed July 2018
[2] https://www.twitter.com, accessed July 2018

2.2 Inspectability and Control in Recommendation Systems

Factors of Inspectability and Control have played an emerging role in areas of intelligent systems and in recent years, micro blogs.

2.2.1 Inspectability. Inspectability across recommendation systems literature is defined as the process of exposing users to the reasoning and data process behind a recommendation. Inspectability of the interface also increases the user's trust in the recommendation system. Authors in [1] designed a hybrid recommendation system which allowed users to understand and control different aspects of the recommendation process instilling factors of inspectability and control. Authors in [14] worked on a modified version of the system built in [1] where they assumed the notion of inspectability be similar to the concept of transparency stated by authors in [29]. Their work concluded that social recommender systems and recommender systems in general do indeed can benefit from facilities that improve the inspectability. Inspectability and control over recommendation algorithms also provide an efficient way of dealing with vast amount of social content. In other previous work, researchers developed a system called *TwitInfo* for visualizing and summarizing important events on Twitter [18]. Furthermore, incorporating explanations and dynamic feedback with recommendation system interfaces have shown to positively impact user perception levels of the recommendation process. While designing our interface, inspectability formed a crucial element of the interface allowing users to browse through the vast amount of relevant social information and understand how items were recommended to them.

2.2.2 Control. Control can be defined as the process of allowing users to interact with different recommendation system options to tweak recommendations. Researchers have implemented different methods of control in their systems which range from rating items to assigning weights to item attributes. In [8], researchers developed *SmallWorlds*, a live Facebook application which had an interactive interface. This was used to control item predictions based upon the underlying data from the Facebook API. Authors in [20] developed a collaborative recommendation system with an interactive interface allowing users to manipulate and tune different options on the interface to generate relevant recommendations. Authors in [1, 14] allowed users to dynamically change and update their preferences during a recommendation process. In our study, we base our interface design to include control to allow users to dynamically modify different system controls (c.f., Section 3).

2.2.3 Impact of Interfaces on Users of Recommender Systems. People's opinions about the items recommended to them and their usability is also directly influenced by the interface of the recommendation system in use. Researchers studied different user interactions with recommender systems and concluded that to design an effective interface, one must consider the following two points: one, what specific user needs are satisfied with the interaction and two, what specific systems features lead to satisfaction of those needs [27]. A more user-centric approach towards evaluating recommendation systems has been suggested in the *ResQue* model by Pearl et al. in [22] which aims to assess the perceived qualities of recommenders such as their usability, usefulness, user's satisfaction and so on.

2.3 Beyond Accuracy

Researchers have stated that accuracy is not always the only criteria which fulfill user satisfaction [10]. Different beyond accuracy metrics have been defined to evaluate recommender systems [7]. Authors in their work [26] show how factors apart from only accuracy can make users more satisfied. Users may also be interested in discovering novel products or in exploring more diverse items. In this sub-section, two main beyond accuracy criteria are discussed: *Novelty* and *Diversity*. Both of these criteria play a critical role for evaluation in this study.

Novelty. This criterion has been defined as "new-original and of a kind which has not been seen before" [31]. More researchers are inclined in the direction that *novelty* is one of the fundamental qualities which can be used to measure a recommendation's effectiveness. *Novel* recommendations are item recommendations that the user was unaware about. Good *novelty* metrics would usually measure how well a recommendation system was able to make a user aware of their previously unknown items [10].

Diversity. This is a concept that has been well studied in the information retrieval literature. It is generally defined as the opposite of similarity. One of the most explored methods for *diversity* is the item-item similarity mostly based on the item content [25]. Authors in [30], state a framework for *novelty* and *diversity* on the basis of three concepts namely: choice, diversity and relevance. In [6], researchers measure *perceived diversity* and overall attractiveness of the recommendation list.

3 MOVIETWEETERS SYSTEM

In this section we look in detail the underlying design of our system, *MovieTweeters*. We define the following two *research goals* for our study: RG1: *Incorporate social information within an existing traditional recommendation system and recommend new and diverse items to users*, RG2: *Study the relationship between beyond accuracy metrics (novelty, diversity) and user satisfaction.*

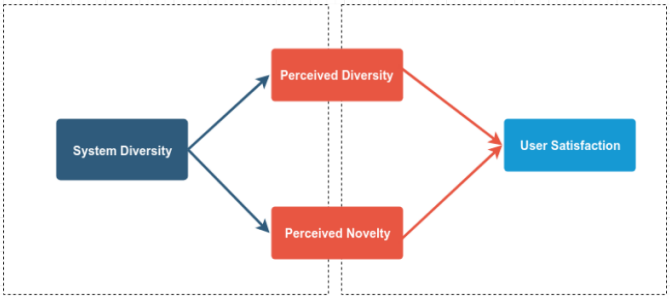

Figure 1: Overview of research goals, studying the effect of offline measures on user perceptions.

In order to study our *research goals*, we define the following research steps in Figure 1. Based on the user-centric framework defined by Knijnenburg et al. [15], we define *system diversity* as an objective system aspect, the perceived user qualities (*perceived novelty* and *perceived diversity*) as subjective system aspects and overall *user satisfaction* as an user experience aspect for our system.

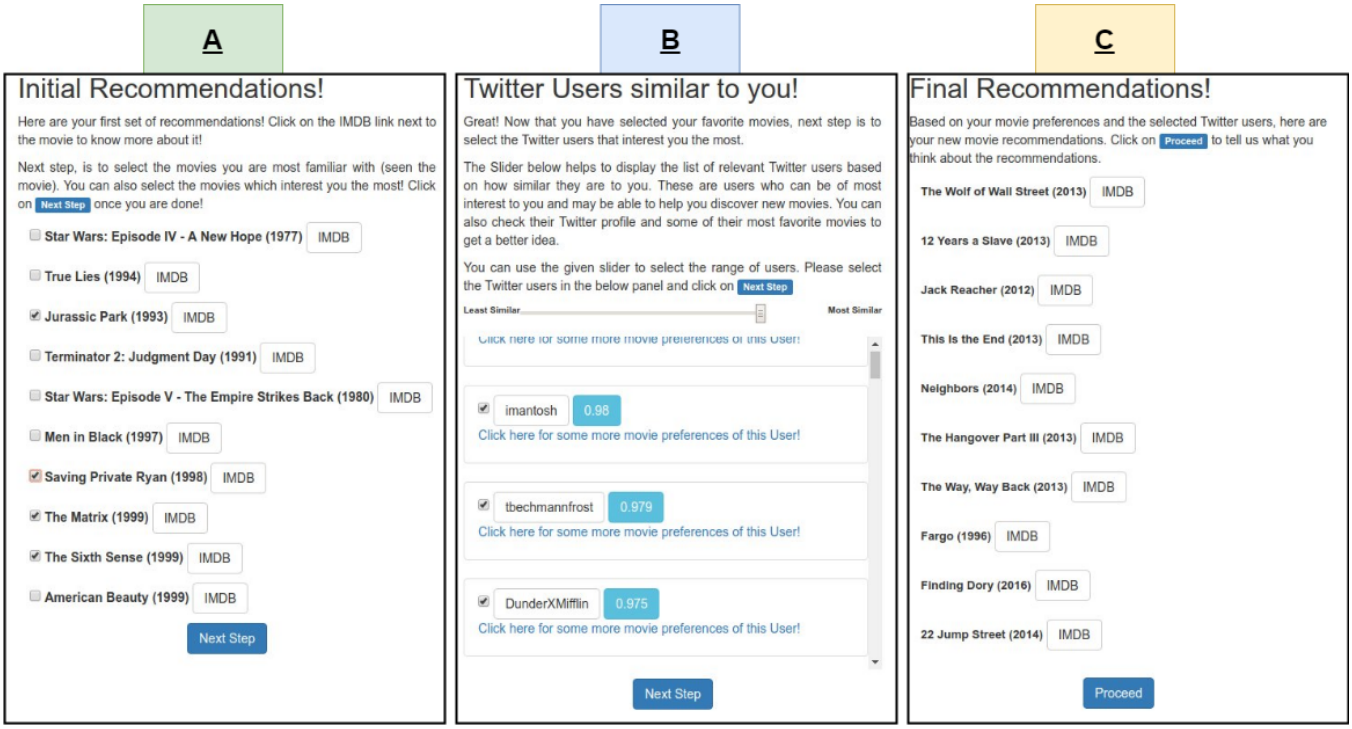

Figure 2: MovieTweeters: A. Initial Recommendations Phase, B. Social Information Phase, C. Revised Recommendations Phase

We first analyze how *system diversity* influences user perception (*perceived diversity* and *perceived novelty*) of the recommended items. Then, we analyze the impact of these perceived qualities on the overall *user satisfaction*.

We designed our system, *MovieTweeters*, a web-based movie recommendation interface, to help us understand the impact of social information on the perceived quality of recommendations and to study the relationships between perceived quality of the recommended items and the overall *user satisfaction*. Figure 2 shows an overview of the system. *MovieTweeters* consists of three main phases namely: Initial Recommendations Phase, Social Information Phase and Revised Recommendations Phase. All three phases were visible to the user when conducting the experiment. Next, we look into these three phases in more detail.

3.1 Initial Recommendations Phase

The initial recommendations phase was involved with the main task of on-boarding new users into our system. We achieved this by first understanding the movie preferences of the new users and generating an initial set of movie recommendations for them.

3.1.1 On-boarding New Users. Recommender systems help suggest items to users they may like based upon the knowledge about the user and the space of available items. However, when new users first enter the system, the system has no information about them. The process of including new users into the system is known as on-boarding. One of the most popular and direct ways to achieve this is to ask the new users to rate an initial set of items, also known as seed items.

To on-board new users in the system and to generate the initial set of movie recommendations, we decided to use the *MovieLens 1M* dataset [9] due to its popularity and being a stable benchmark dataset. First step, was to select the initial seed items. One common selection strategy is to use the *popularity* measure while determining the seed items. In this, the items are ranked in a decreasing order of the number of ratings. The*Movielens 1M* dataset suffers from a long tail distribution problem [21]. This essentially means there are some movies in the dataset which have been very frequently rated (the popular movies). Some methods to deal with this phenomenon was explained using diffusion theory [11] and graph-based approaches [32].

Using the *popularity* strategy to select seed items would end up selecting items which are most popular (highly rated) in the dataset and would ignore the unpopular or new ones. This could create a bias where only popular (highly rated) movies are recommended. In order to avoid this bias, we based our approach to select items based on 2 criteria: *popularity* (number of times rated) and *ratings* of the movies (as defined by the authors in [23]). We calculated a seed score for each movie in the dataset using the following approach:

$$seedscore = \log(popularity) \times \sigma^2(ratings) \quad (1)$$

To calculate the seed score (Equation 1), we first took the *logarithm (base 10)* of the *popularity*. In the second half of the formula, we consider the *variance* of the *ratings*. This gives us a measure of how diverse the ratings have been for a particular movie.

3.1.2 Recommendation Algorithm. In our system, we used *Item-Item* Collaborative Filtering algorithm [24] to generate the initial

set of movie recommendations for the new users. As the main focus of this study is the role and impact of recommendation interfaces, we decided to use a Collaborative Filtering algorithm to make the initial set of movie recommendations because of their popularity and simplicity of use and implementation. We used the GraphLab toolkit [16] to implement it in our system.

3.1.3 Process Flow. After on-boarding new users into our system, a set of 10 initial movie recommendations were generated for them (as shown in Initial Recommendations Phase in Figure 2). We refer to these first list of top-10 recommendations as Recommendation List 1 (**RL1**).

3.2 Social Information Phase

The social information phase was mainly responsible for incorporating a relevant social information dataset into our system.

3.2.1 Social Information Dataset. We used *MovieTweetings* [5] as our social information source. *MovieTweetings* comprises of IMDb ratings expressed by Twitter users who have connected their IMDb accounts to their Twitter account.

3.2.2 Process Flow. The next task for the system user was to select the most *relevant movies* out of the initial recommendation list (**RL1**). *Relevant movies* here is defined as the movies which the system user has already watched (consumed) or the movies which seem the most interesting to him/her. After the *relevant* movies were selected by the system user, the next step was to retrieve the relevant Twitter users from our pre-processed *MovieTweetings* dataset and display them. This was a two step process: First, all distinct Twitter users who rated at least one of the selected *relevant* movies were first retrieved. Second step was calculating how similar these retrieved Twitter users were to the system user. In order to perform this task, we first retrieved all movies rated by each Twitter user individually. For the system user, we considered all the movies he/she rated from the initial seed items. We ran the *cosine similarity* to measure the relevant similarity between the genre distribution of movies consumed by the system user and each of the Twitter User. This produced a *similarity score* which denoted how similar was a given Twitter user to the system user.

We displayed the Twitter Users in the decreasing order of their *cosine similarity scores* (as shown in Social Information Phase in Figure 2). A slider was also included in the system which allowed the system users to segregate Twitter users based on their *similarity score*. The system users had to select at least one preferred Twitter user. There was no limit on the maximum number of users selected.

3.3 Revised Recommendations Phase

The revised recommendations phase was responsible for generating a revised list of movie recommendations. After the most preferred Twitter users were selected by the system user, the next step was to retrieve a list of all the movies rated by these Twitter users from our pre-processed *MovieTweetings* dataset. We calculated the *local popularity* score of each movie in the list of retrieved movies. *Local popularity* is basically the number of occurrences of the movie in the list. We then ordered this list of movies in a descending order based on their *local popularity* scores. We refer to this movie list as the *relevant movie list*. Figure 3 shows a description of this process.

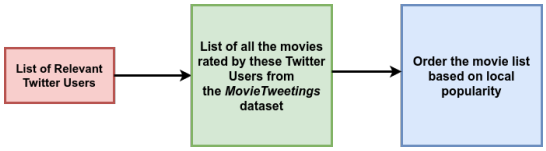

Figure 3: Selecting and ordering of retrieved relevant movies

3.3.1 Maximal Marginal Relevance. It is one of the diversity-based re-ranking methods used for reordering ranked list of documents [3]. Following its success in the field of text retrieval and summarization, we tweaked the Maximal Marginal Relevance method and apply it to our *relevant movie list*. It was calculated using a weighted linear combination of *relevance* and *diversity* [3]:

$$MMR \triangleq \max_{Di \in R/S} [\lambda(Rel(D_i, Q)) - (1-\lambda)(\max_{D_j \in S}(1 - Sim(D_i, D_j))] \quad (2)$$

In Equation 2, R refers to the ranked list of movies in the *relevant movie list*. R/S is the set difference, i.e, the set of unselected movies from R. S refers to the movies which are already retrieved. The first half of the equation, $\lambda(Rel(D_i, Q))$ addresses the *relevance* aspect of the equation and is calculated by comparing how similar is D_i is to Q. D_i refers to a movie in the *relevant movie list*. Q for a given system user is calculated by considering the movies (from the Initial Recommendations Phase) which were selected by him/her. This similarity was calculated by comparing the genres of the list of movies in Q with D_i using *cosine similarity*. This gave a *relevance* score which constituted the first half of the equation.

The second half of the equation $(1-\lambda)(\max_{D_j \in S}(1-Sim(D_i, D_j))$ addresses the *diversity* aspect of the MMR equation. S refers to the movies which are already selected in the re-ranked MMR List. Here, *cosine similarity* was used to calculate how similar D_i was with D_j based on their genre details. This constituted a score for *diversity* which formed the second half of the equation. These two scores are combined together to form the final MMR score.

Based on the MMR scores, a revised list of movie recommendations, both relevant to the system user and diverse according to their tastes, was generated and presented to them (as shown in Revised Recommendations Phase in Figure 2). The size of the MMR list was set to be 10 (*top-10* recommendations). We refer to this revised list of *top-10* recommendations as Recommendation List 2 (**RL2**).

4 EXPERIMENT

We evaluated our system *MovieTweeters* using both offline and online metrics. We seek to answer the following: First, the impact of social information (with an interactive interface) on the quality of recommendations. Second, the relationship between the quality of the recommendations and user satisfaction. The offline evaluation metrics which we define later in this section, help us analyze how an additional input of social information with a traditional recommendation system affects *system diversity*. The online evaluations help us analyze the user perceptions (*perceived novelty & perceived diversity*) of the recommendations and understand their satisfaction levels.

4.1 Variable Description

We describe all the dependent and independent variables which are evaluated during the course of the experiment and also form a part of our hypotheses. There is one main independent variable in our experiment:

- **System Diversity:** In the context of this study, we define *System Diversity* for the recommended items (movies) in terms of how different they are in terms of genre.

Based on the independent variable defined above, we now define our two dependent variables whose effects will be evaluated and tested during the course of the experiment.

- **Perceived Novelty:** The extent to which users receive "new" movie recommendations. Here, we evaluate whether users are able to come across movies which they have not seen before. It is derived from the *ResQue* framework developed by Pearl et al. in [22] which accesses the quality of the recommended items.
- **Perceived Diversity:** The extent to which users felt that the recommended items were *diverse* to them. It is defined from the "Perceived System Qualities" stated in [22] by Pearl et al.

4.2 Hypotheses

Our system *MovieTweeters* generated two recommendation lists of movies for the users; one before **(RL1)** and one after **(RL2)** the participant's interaction with their relevant social information. We hypothesized that our system will help the users discover more novel and diverse content. Following are our hypotheses:

Hypothesis H_1: *System Diversity* has a correlation with the *Perceived Diversity* of the participants on the two recommendation list items.

Hypothesis H_2: *System Diversity* has a correlation with the *Perceived Novelty* of the participants on the two recommendation list items.

Hypothesis H_3: *Perceived Novelty* of the users increases between the two lists.

Hypothesis H_4: As the *Perceived Novelty* of participants increases, their *User Satisfaction* increases as well.

Hypothesis H_5: As the *Perceived Diversity* of users increases, their *User Satisfaction* increases as well.

4.3 Materials

We used two datasets to make movie recommendations in our system.

4.3.1 MovieLens 1M Dataset. The first one, to generate the initial list of recommendations (**RL1**), we used the *MovieLens 1M* dataset. Our pre-processing steps included adding relevant IMDb IDs to the movies, making sure all movies had their relevant genre information present, removal of irrelevant fields such as time-stamp of the ratings. Our pre-processed *MovieLens 1M* dataset had 964712 ratings from 6040 users for 2835 movies.

4.3.2 MovieTweetings Dataset. We used *MovieTweetings* as our social information source. We used this dataset to make our revised set of movie recommendations (**RL2**). Our pre-processing steps included removal or irrelevant fields such as time-stamp of the ratings, retrieving the Twitter IDs of the users in the dataset, removal of movies with no relevant genre information present. Our pre-processed *MovieTweetings* dataset had 606767 ratings from 45871 users for 27093 movies.

4.4 Experimental design

Keeping in mind the hypotheses stating the impact of *system diversity* on user perception and the relationship between user perception of the recommendations with their satisfaction (defined in Section 4.2), we study and analyze the impact of different system variables on the perceived quality attributes.

4.4.1 Evaluation Metrics. We study the impact of *system diversity* on *perceived novelty* and *perceived diversity*.

- **System Diversity:** We used *Intra List Diversity* [2] to calculate the *system diversity* of the two generated recommendation lists (**RL1** & **RL2**) in our system. For our study, we define it as the following:

$$ILD = \frac{\sum_{i=1}^{n}\sum_{j=i}^{n}(1 - sim(c_i, c_j))}{n * (n-1)/2} \tag{3}$$

  where $c_1, c_1 ... c_n$ are items in a given list and n refers to the total number of items in the list. We used the Cosine similarity measure to calculate the distance between the items.
- **Perceived Novelty:** For our study, we defined *Novelty* as movies which the user has never seen before. We measured *Perceived Novelty* with the following two processes:
  - **Perceived Novelty of the Recommendation Lists:** Participants selected novel items from both recommendation lists (**RL1** and **RL2**).
  - **Perceived Novelty Questionnaire:** We asked the participants to answer a set of three questions which helped us understand their level of *perceived novelty* across the two lists:
    - **Q 1:** *The movies recommended in Recommendation List 1 were interesting to me.*
    - **Q 2:** *The movies recommended in Recommendation List 2 were interesting to me.*
    - **Q 3:** *The Twitter Users helped me obtain novel movie recommendations and improved the overall recommendation process.*
- **Perceived Diversity:** We asked the participants to answer four questions about their level of *perceived diversity* between the two lists:
  - **Q 1:** *The list of movies in Recommendation List 2 vary from the list of movies in Recommendation List 1.*
  - **Q 2:** *Most of the movies in Recommendation List 2 belong to similar genres as Recommendation List 1.*
  - **Q 3:** *The movies recommended to me in Recommendation List 2 are diverse.*

- **Q 4:** *Selecting the relevant Twitter Users helped me obtain diverse movie recommendations and improved the overall recommendation process.*

We also analyze the impact of *perceived novelty* and *perceived diversity* on overall *user satisfaction*.

- **User Satisfaction:** We define *User Satisfaction* not only in terms of how satisfied they are with the quality of the recommendations but also their experience with the inspectability, control and overall interface aspects of our system. We asked the participants to answer a set of six questions which helped us understand their overall *user satisfaction*:
  - **Q 1:** *The recommendation system provided me with good movie suggestions.*
  - **Q 2:** *The recommendation system helped me discover new movies.*
  - **Q 3:** *The movies recommended to me are diverse.*
  - **Q 4:** *The recommendation system made me more confident about my selection/decision.*
  - **Q 5:** *I am convinced I will like the movies recommended to me.*
  - **Q 6:** *Overall, I am satisfied with the recommendation system and the interface.*

4.4.2 MMR Value. The value of λ is used to adjust the *relevance* and the *diversity* scores to emphasize between relevance & diversity. In our system, we made the design decision to have an equal balance between the *relevance* and the *diversity* aspects of the revised list of recommended movies. Hence, the value of λ was set to **0.5**.

4.4.3 Procedure and Tasks. As seen in Section 3, we divide our system into three phases:

- **Initial Recommendations Phase:**
  (1) Participants were asked to provide basic demographic information (gender, age, movie consumption details) and to rate at least 20 movies out of 40 movies (initial seed).
- **Social Information Phase:**
  (1) Participants were asked to select their relevant list of movies from the initial recommendation list (**RL1**) and select their most preferred Twitter users
- **Revised Recommendations Phase:**
  (1) Participants were shown the revised list of recommendations based on their selections (**RL2**).

After completion of the main experiment, they had to answer different post evaluation questionnaires which evaluated different aspects of both the recommendation lists and also their user satisfaction. Responses to the post evaluation questionnaires were collected in the form of a Likert scale (1-5). We followed a within-subjects experimental design where all the participants were exposed to both their recommendation lists (**RL1** and **RL2**) and were made to compare the two lists.

5 RESULTS

In this section, we discuss the results we found regarding the impact of *system diversity* of the perceived quality of recommendations, and also the relationship between the perceived quality of recommendations and *user satisfaction*.

5.1 Participants

The experiment was held in a controlled setting with 23 participants. Each interview lasted 15-25 minutes. The participants were mainly Master students at a university with a varied educational background. We had an equal gender distribution with 52.2% female and 47.8% male participants. Most of the participants were between the ages of 25-34 (60.9%).

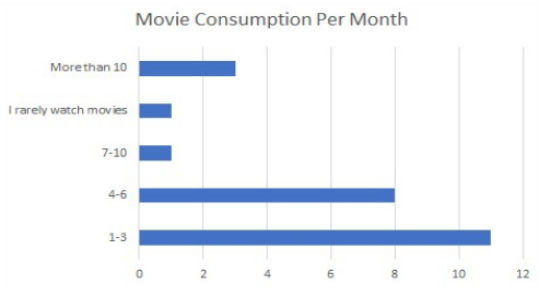

Figure 4: Movie Consumption Behavior

Figure 4 gives us an overview of their movie consumption patterns with most participants in the range of 1-6 movies consumed per month.

5.2 Offline Evaluation

System Diversity. Using the offline evaluation metric defined in Section 4, we calculated the **system diversity** (*Intra List Diversity*) of both the generated recommendation lists (**RL1** and **RL2**). Overall, we found a significant increase in the system diversity across the two recommendation lists (Figure 5).

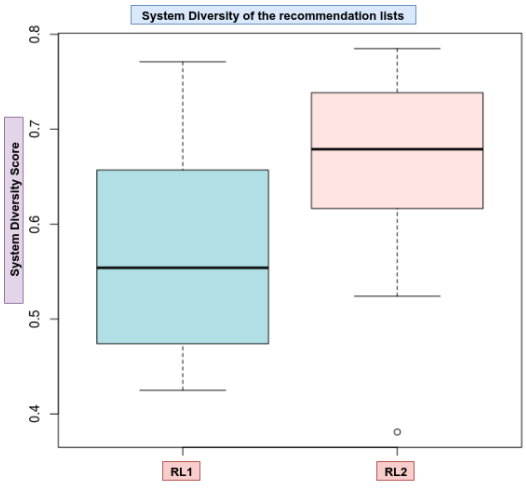

Figure 5: System Diversity Scores, before (RL1) and after (RL2) the interaction.

5.3 Online Evaluation

5.3.1 Hypothesis 1: System Diversity and Perceived Diversity. To validate our Hypothesis $\mathbf{H_1}$, we ran the Spearman's Rank-Order

Correlation test to compare the impact of the change in *System Diversity* between the recommendation lists on the *Perceived Diversity* of the participants on the recommended items from both the recommendation lists (**RL1** and **RL2**). Spearman's correlation coefficient ρ measures the strength and direction of the correlation between two associated variables.

We observed that the ρ **value** is **-0.54** and it is significant ($\mathbf{p} = \mathbf{0.007}, \mathbf{p} < \mathbf{0.05}$), this demonstrates a strong negative correlation between the *System Diversity* and the *Perceived Diversity* of the recommended items in both the recommendation lists. This rejects the null hypothesis and our alternative hypothesis ($\mathbf{H_1}$) is accepted.

5.3.2 Hypothesis 2: System Diversity and Perceived Novelty. To validate our Hypothesis $\mathbf{H_2}$, we ran the Spearman's Rank-Order Correlation test to compare the impact of the change in *System Diversity* between the recommendation lists on the *Perceived Novelty* of the participants on the recommended items from both the recommendation lists (**RL1** and **RL2**).

We observed that the ρ **value** is **-0.42** and it is significant ($\mathbf{p} = \mathbf{0.04}, \mathbf{p} < \mathbf{0.05}$), demonstrating a strong negative correlation between the *System Diversity* and the *Perceived Novelty* of the recommended items in both the recommendation lists. This rejects the null hypothesis and our alternative hypothesis ($\mathbf{H_2}$) is accepted.

5.3.3 Hypothesis 3: Perceived Novelty Increases. To validate our Hypothesis $\mathbf{H_3}$, we ran the Wilcoxon Signed Rank Test to compare the difference (before and after) between the two recommendation lists (**RL1** (before) and **RL2** (after)) in terms of novel items.

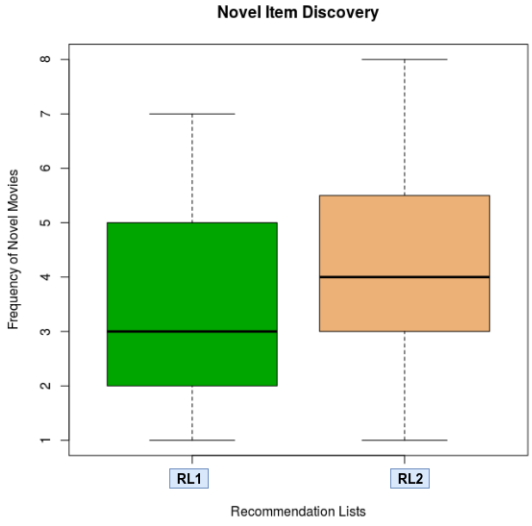

Figure 6: Number of items marked as novel, before (RL1) and after (RL2) the interaction.

We observe that there is statistical significance in the *Perceived Novelty* of the participants after interacting with the relevant Twitter information ($\mathbf{p} = \mathbf{0.0009}$, $\mathbf{p} < \mathbf{0.01}$). Figure 6 shows the frequency of *novel* movies for both the lists (**RL1** & **RL2**). Analyzing all the data statistics and the relevant information, we infer that the participants did indeed find more novel items after interaction with the relevant social information. This rejects the null hypothesis and our alternative hypothesis ($\mathbf{H_3}$) is accepted.

5.3.4 Hypothesis 4: Perceived Novelty and Satisfaction. To validate our Hypothesis $\mathbf{H_4}$, we ran the Spearman's Rank-Order Correlation test to study the impact of *Perceived Novelty* of the participants on their overall *User Satisfaction*.

We observed that the ρ **value** is **0.70** and it is significant ($\mathbf{p} = \mathbf{0.00016}, \mathbf{p} < \mathbf{0.05}$). The ρ value shows that there is a strong positive correlation between the *Perceived Novelty* and *overall User Satisfaction* of the participants. This rejects the null hypothesis and our alternative hypothesis ($\mathbf{H_4}$) is accepted.

5.3.5 Hypothesis 5: Perceived Diversity and Satisfaction. To validate our Hypothesis $\mathbf{H_5}$, we ran the Spearman's Rank-Order Correlation test to study the impact of *Perceived Diversity* of the participants on their overall *User Satisfaction*. We observed that the ρ **value** is **0.58** and it is significant ($\mathbf{p} = \mathbf{0.003}, \mathbf{p} < \mathbf{0.05}$), demonstrating a strong positive correlation between the *Perceived Diversity* and *overall User Satisfaction* of the participants. This rejects the null hypothesis and our alternative hypothesis ($\mathbf{H_5}$) is accepted.

6 DISCUSSION AND LIMITATIONS

In this section we discuss our initial results from the experiment and also the limitations. According to our research goals defined in Section 3, we wanted to build an interactive interface that could assist users to discover more *novel* content. Our initial results (for Hypothesis H_3) suggested that we were successful in this aspect and our interface indeed helped users discover more *novel* items.

While analyzing the impact of *system diversity* (of the recommendations) on the perceived measures of quality (*perceived diversity* and *perceived novelty*), we found some surprising results (Hypotheses H_1 and H_2). Notably as *system diversity* increased across the two lists, the *perceived diversity* and *perceived novelty* of the users decreased. The decrease in the *perceived diversity* of the users could be attributed to the following factors. Our MMR formula ($\lambda = 0.5$ balanced out the revised recommendation list in terms of *relevance* and *diversity*. This could impact the perception of the participants in a way where they were actually focused on checking the *relevance* aspect of the recommendations. In a post-hoc analysis, we studied the impact of *popularity* of the *diversity perception* of the users. *Popularity* for a recommendation list was calculated by taking the average of the top 3 IMDb movie ratings for that list. We observed that the popularity scores actually decreased across the lists. We compared this to the *perceived diversity* of the participants and found that as the *popularity* decreased across the lists, the *perceived diversity* decreased as well. We state that popularity also played a role in affecting the *perceived diversity* of the participants. Diversification (increase in the *system diversity*) led to less popular movies which made the participants perceive them as less *diverse*.

We found a positive correlation between users *perceived novelty* and *perceived diversity* on overall *user satisfaction* (Hypotheses H_4 and H_5). Users who perceived that they discovered more novel and diverse items reported increased levels of *satisfaction*.

6.1 Limitations

We identify four main limitations in our study:

- An experiment comparing different recommendation algorithms could prove helpful to understand the impact these different recommendation algorithms have on perceived quality of the recommendations and on overall user satisfaction.
- Multiple experiments with different MMR (λ) values could help us understand its impact on the revised recommendation list and ultimately on the perception of the users and their overall user satisfaction.
- The inclusion of other movies specific features (e.g., features such as box office revenues, year/era of release) could provide more insight into how two users are correlated which would also impact the final recommendation list.
- Recent research has studied the impact of personality on the diversity needs of users [4, 17]. Our study did not consider the effects of individualistic traits such as personality.

Overall, our interface proved that the incorporation of relevant social information with an interactive interface does indeed help users discover more *novel* items. It also provided valuable insight into the relationship between how users perceive the recommended items and their overall satisfaction.

7 CONCLUSION

In this paper, we introduced and evaluated a novel interface, *MovieTweeters*. It is a movie recommender system which combines social information with a traditional recommendation algorithm. This allows us to generate recommendations that are both current (since the social information is constantly updating), and novel. We conducted offline and online evaluations to test our interface. We found that incorporation of social information with an interactive interface can indeed help users discover more *novel* content. Also, we observed that users who *perceived* that they discovered more *novel* and *diverse* items also reported increased levels of *user satisfaction*. Even though we successfully were able to increase the *system diversity* of the recommendations, it had a negative correlation with users perception of *novelty* and *diversity* of the items.

In future work, inclusion of different recommendation algorithms along with varying values of MMR will be studied. We believe it could have a significant impact on how users perceive their recommendations, and also on their overall satisfaction.

REFERENCES

[1] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: a visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on Recommender systems*. ACM, 35–42.

[2] Keith Bradley and Barry Smyth. 2001. Improving recommendation diversity. In *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland*. Citeseer, 85–94.

[3] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 335–336.

[4] Li Chen, Wen Wu, and Liang He. 2013. How personality influences users' needs for recommendation diversity?. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM, 829–834.

[5] Simon Dooms, Toon De Pessemier, and Luc Martens. 2013. Movietweetings: a movie rating dataset collected from twitter. In *Workshop on Crowdsourcing and human computation for recommender systems, CrowdRec at RecSys*, Vol. 2013. 43.

[6] Bruce Ferwerda, Mark P Graus, Andreu Vall, Marko Tkalcic, and Markus Schedl. 2017. How item discovery enabled by diversity leads to increased recommendation list attractiveness. In *Proceedings of the Symposium on Applied Computing*. ACM, 1693–1696.

[7] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 257–260.

[8] Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Christopher Hall, and Tobias Höllerer. 2010. Smallworlds: visualizing social recommendations. In *Computer Graphics Forum*, Vol. 29. Wiley Online Library, 833–842.

[9] F Maxwell Harper and Joseph A Konstan. 2016. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4 (2016), 19.

[10] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004), 5–53.

[11] Masayuki Ishikawa, Peter Geczy, Noriaki Izumi, and Takahira Yamaguchi. 2008. Long tail recommender utilizing information diffusion theory. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 785–788.

[12] Henry Kautz, Bart Selman, and Mehul Shah. 1997. Referral Web: combining social networks and collaborative filtering. *Commun. ACM* 40, 3 (1997), 63–65.

[13] Irwin King, Michael R Lyu, and Hao Ma. 2010. Introduction to social recommendation. In *Proceedings of the 19th international conference on World wide web*. ACM, 1355–1356.

[14] Bart P Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems*. ACM, 43–50.

[15] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 441–504.

[16] Yucheng Low, Danny Bickson, Joseph Gonzalez, Carlos Guestrin, Aapo Kyrola, and Joseph M Hellerstein. 2012. Distributed GraphLab: a framework for machine learning and data mining in the cloud. *Proceedings of the VLDB Endowment* 5, 8 (2012), 716–727.

[17] Feng Lu and Nava Tintarev. 2018. A Diversity Adjusting Strategy with Personality for Music Recommendation. In *IntRS@ RecSys*.

[18] Adam Marcus, Michael S Bernstein, Osama Badar, David R Karger, Samuel Madden, and Robert C Miller. 2011. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 227–236.

[19] Paolo Massa and Paolo Avesani. 2007. Trust-aware recommender systems. In *Proceedings of the 2007 ACM conference on Recommender systems*. ACM, 17–24.

[20] John O'Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. 2008. PeerChooser: visual interactive recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1085–1088.

[21] Yoon-Joo Park and Alexander Tuzhilin. 2008. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM conference on Recommender systems*. ACM, 11–18.

[22] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 157–164.

[23] Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K Lam, Sean M McNee, Joseph A Konstan, and John Riedl. 2002. Getting to know you: learning new user preferences in recommender systems. In *Proceedings of the 7th international conference on Intelligent user interfaces*. ACM, 127–134.

[24] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. ACM, 285–295.

[25] Barry Smyth and Paul McClave. 2001. Similarity vs. diversity. *Case-Based Reasoning Research and Development* (2001), 347–361.

[26] Kirsten Swearingen and Rashmi Sinha. 2001. Beyond algorithms: An HCI perspective on recommender systems. In *ACM SIGIR 2001 Workshop on Recommender Systems*, Vol. 13. 1–11.

[27] Kirsten Swearingen and Rashmi Sinha. 2002. Interaction design for recommender systems. In *Designing Interactive Systems*, Vol. 6. 312–334.

[28] Jiliang Tang, Xia Hu, and Huan Liu. 2013. Social recommendation: a review. *Social Network Analysis and Mining* 3, 4 (2013), 1113–1133.

[29] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender Systems Handbook*. Springer, 353–382.

[30] Saúl Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 109–116.

[31] Liang Zhang. 2013. The Definition of Novelty in Recommendation System. *Journal of Engineering Science & Technology Review* 6, 3 (2013).

[32] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* 107, 10 (2010), 4511–4515.