Enhancing User Experience with Recommender Systems Beyond Prediction Accuracies

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Tien Tran Tu Quynh Nguyen

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Joseph A. Konstan and Loren Terveen

August, 2016

Acknowledgements

I would like to thank:

- My advisors, Joseph Konstan, Loren Terveen and John Riedl [1] for their continuous support, patience, criticism and advice during my graduate study.

- Professor Daniel Keefe and Professor Adam Rothman for their support and advice as my committee members.

- My colleagues, Daniel Kluver, Vikas Kumar, Pik-Mai Hui, Raghav Karumur, Aaron Halfaker, Bowen Yu, Steven Chang, Max Harper, Michael Ekstrand and others who have supported, discussed and collaborated with me on several research projects.

- GroupLens for creating and maintaining a friendship and intellectual-curiosity environment. My English professor Colleen Meyers for her helps and chats during my years of studying.

- My family, especially my wife Mary Duyen Nguyen, for her continuous support and encouragement over the last five years. I am also thankful for my parents for their support during my years of studying.

[1] Deceased July, 15 2013

Dedication

To the Orion (Mary + Alex) and Tu Quynh Families.

Abstract

In this dissertation, we examine to improve the user experience with recommender systems beyond prediction accuracy. We focus on the following aspects of the user experience. In chapter 3 we examine if a recommender system exposes users to less diverse contents over time. In chapter 4 we look at the relationships between user personality and user preferences for recommendation diversity, popularity, and serendipity. In chapter 5 we investigate the relations between the self-reported user satisfaction and the three recommendation properties with the inferred user recommendation consumption. In chapter 6 we look at four different rating interfaces and evaluated how these interfaces affected the user rating experience.

We find that over time a recommender system exposes users to less-diverse contents, and that users rate less-diverse items. However, users who took recommendations were exposed to more diverse recommendations than those who did not. Furthermore, users with different personalities have different preferences for recommendation diversity, popularity, and serendipity (e.g. some users prefer more diverse recommendations, while others prefer similar ones). We also find that user satisfaction with recommendation popularity and serendipity measured with survey questions strongly relate to user recommendation consumption inferred with logged data. We then propose a way to get better signals about user preferences and help users rate items in the recommendation systems more consistently. That is providing exemplars to users at the time they rate the items improved the consistency of users' ratings.

Our results suggest several ways recommender system practitioners and researchers can enrich the user experience. For example, by integrating users' personality into recommendation frameworks, we can help recommender systems deliver recommendations with the preferred levels of diversity, popularity, and serendipity to individual users. We can also facilitate the rating process by integrating a set of proven rating-support techniques into the systems' interfaces.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Many people are relying on recommender systems such as Amazon and Netflix to make decisions about things to consume. The perceived benefits of recommender systems are in part due to their ability to predict how much users would enjoy unknown items based on some limited data provided by users.

Indeed, much of early work in recommender systems focused on improving this particular ability of recommender systems. The focus has resulted in many well-known recommendation algorithms, such as user-user collaborative filtering (CF), item-item CF [2], or more sophisticated approaches like matrix factorization [3]. These algorithms are used widely in well-known systems like Amazon or Netflix.

Recommender system researchers, however, have argued that many other factors, besides prediction accuracy, affect the user experience with recommender systems. Laughlin et al. [4] pointed out that accuracy metrics are not capable of capturing user experience because the metrics are designed to assess the prediction accuracy of an individual recommendation. Konstan et al. [5] suggested that accurately predicted ratings are not good to evaluate how well a system recommends valuable items previously unknown to users. Moreover, McNee et al. [6] argued that we should evaluate the quality of recommendations as a list, not as individual items in the list. Thus, as Hayes et al. [7] asserted, research in recommender systems needs to go beyond pure accuracy and toward real user experience.

The experience of users with recommender systems starts even before they receive recommendations. It begins when when users sign up to use a system. McNee et al. [8] showed that that letting users choose items to rate, instead of providing them to users, during the sign-up process increases user loyalty to the systems. This is because users thought the systems could learn their profile quickly, despite the fact that users took a longer time to sign-up than usual. Or Drenner et al. [9] showed that asking users to apply tags to movies when they first joined a recommender system shaped their subsequent behaviors in the system.

User experience with recommender system does not end after signing up. After sign-up, a user life-cycle with recommender systems consists of two phases. At first, users test if recommender systems work for them. Then, they take advantage of recommender systems for valuable recommendations [10,11]. Thus, what users receive during the two phases can affect user experience, especially their trust in recommender systems.

Many factors can affect the experience during user life-cycles with recommender systems. For example, Tintarev et al. [12] argue that transparency builds user trust with the systems, and lets users take actions if the users see systems' mistakes. Swearingen et al. [13] provide evidence that users perceived recommender systems are effective when the systems inspire trust, are transparent about the recommendations, and *"points users to new and not-yet-experienced items"*.

In this dissertation, we examine ways to improve the user experience with recommender systems beyond prediction accuracy. Next, we define our research goals.

1.2 Research Goals

We focus on several aspects of user experience. These aspects are broad - from the long-term effect of a recommender system on users to user satisfaction with the three fundamental properties of recommendations, to rating interfaces. Because these factors appear in the two principal stages of users' interactions with a recommender system: rating items and receiving recommendations, these factors are important for researchers and practitioners to consider when they evaluate the experiences of their users with their recommender systems.

We dedicate each chapter to describe our study for each user experience factor and explain the implications of the findings related to the factor. We first present our study on whether a recommender system exposes users to a less diverse content over time (chapter 3). We then discuss the possibility of improving user experience by integrating user personality in the process of generating recommendations (chapter 4). Next, we then discuss the relations between user experience measurements via survey questionnaires and logged data analyses (chapter 5). Finally, we discuss how to improve user experience during their rating process (chapter 6). We conclude the dissertation by summarizing our findings, our limitations, and putting forward ideas that can build upon our results.

1.3 Research Context

1.3.1 MovieLens Platform

We carry all of our research projects on a live movie recommendation called MovieLens[1]and its dataset. MovieLens is a movie recommender system that has been in continous use since 1997 with more than 200,000 users providing more than 20 million ratings. Users provide their ratings on a rating scale from 0.5 star to 5 stars with 0.5 star increments.

[1] www.movielens.org

We conduct our research projects in chapters 3 and 6 with the third version of MovieLens (figure 1.2). In November 2014, MovieLens was upgraded to the fourth version (the latest version as of the time of writing, figure 1.3). We conduct research projects in chapters 4 and 5 with data collected on the latest version.

1.3.2 Tag-genome

In the research projects described in chapters 3, 5, 4, and 6, we use the tag-genome information space[2] to approximate the content of movies. Subsequently, we use this approximated movie content to compute the diversity, and serendipity of lists of movie recommendations.

The *tag-genome* [14] is an information space containing a set M of movies, and a set T of tags. The set T consists of tags that are highly descriptive about movies. How well tag t describes movie m is expressed via the relevance score $rel(t, m)$. The relevance score $rel(t, m)$ takes on values from 1 (does not describe the movie m at all) to 5 (strongly describes the movie m)[3]. Each movie m_i is represented as a vector of size $|T|$ where entry i, j is the relevance of tag j to movie i. Figure 1.1 visualizes the tag genome information space.

Vig et al. [14] introduced the tag-genome in their work in 2012. To compute the relevance score for a pair of a movie and a tag, they presented a systematic machine learning approach consisting of the following steps. In the first step, they selected a set of tags that are large enough to capture information about movies. Thus, relevance scores are the predicted scores produced by their machine learning models for these tags and movies. In the second step, they trained the models to do the prediction tasks. To train the models, they extracted features about movies. These extracted features are user-generated content about movies (such as reviews, comments, blogs, ratings, or tags), and the meta-data about movies (genres, title, artist, or released date), and the content of an item (such as movie's

[2] The tag-genome data set is available at http://grouplens.org/datasets/movielens/

[3] Vig et al. [14] originally proposed the range to be from 0 to 1. However, after several revisions, as of 09/26/2013 MovieLens uses the range of 1 to 5.

audio file). They also collected training data based on the inputs of MovieLens.

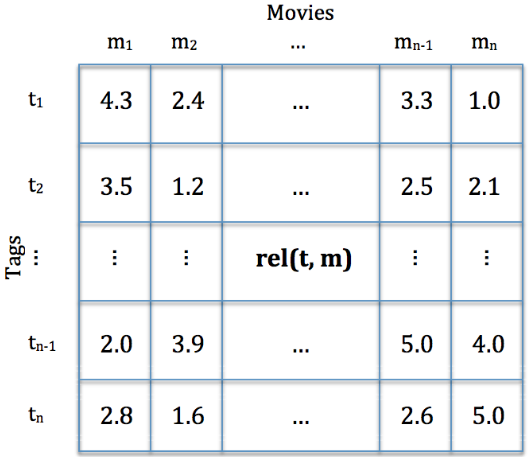
1.3.3 User Personality

To answer the research questions in chapter 4, we need a dataset about user personality and their ratings. However, it has been pointed out that such a data set is not freely available for research in recommender systems [15]. Thus, we run an experiment on MovieLens for five months (from May 2015 till October 2015) to collect the information about user personality and their experience with MovieLens recommender system.

In the experiment, we measure user personality based on the Big Five Personality Traits established in personality psychology research [16, 17]. These five big personality traits are *agreelableness, conscientiousness, extreversion, neuroticism, and openness*. To measure user personality trait, we use the 10 questions proposed by Gosling et al. [18]. We will discuss our process of collecting user personality in details in chapter 4.

1.4 Outline

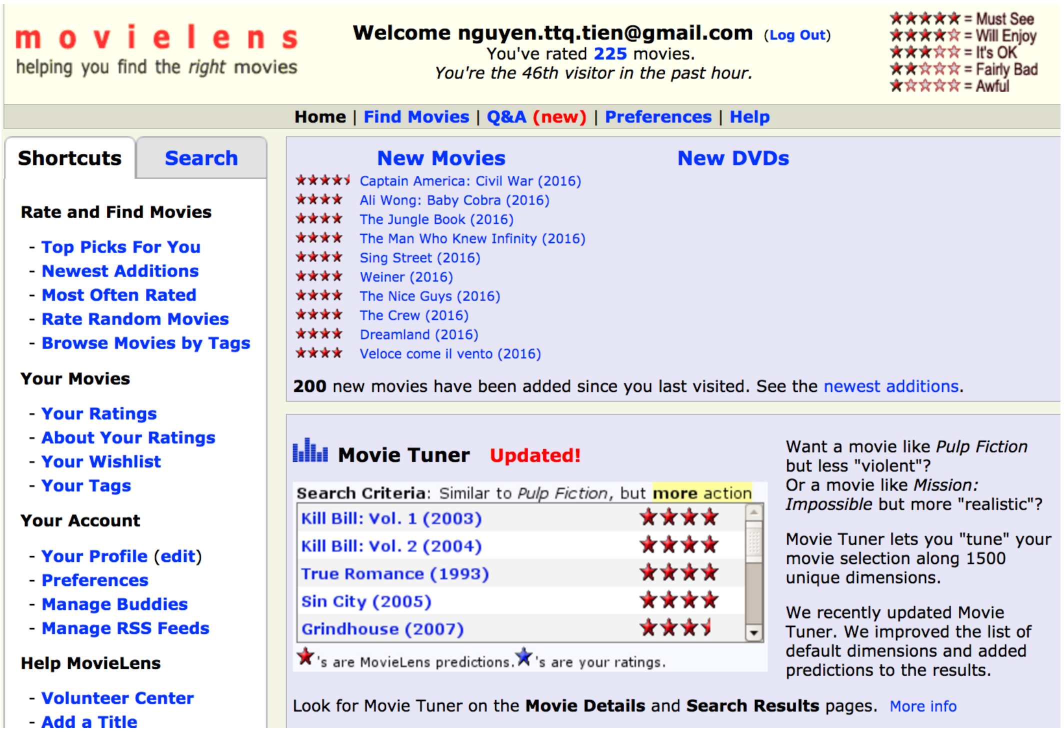
The next chapter presents prior work related to user experience with recommender systems. The following chapters present our research corresponding to the goals mentioned above. In chapter 7, we summarize our contributions of this dissertation and conclude with a discussion of the most promising areas for future work.



Movies

Tags	m_1	m_2	...	m_{n-1}	m_n
t_1	4.3	2.4	...	3.3	1.0
t_2	3.5	1.2	...	2.5	2.1
⋮	⋮	⋮	**rel(t, m)**	⋮	⋮
t_{n-1}	2.0	3.9	...	5.0	4.0
t_n	2.8	1.6	...	2.6	5.0

Figure 1.1: The visualization of tag-genome. This figure is reproduced based on the original figure in Vig et al. [14].



movielens
helping you find the *right* movies

Welcome nguyen.ttq.tien@gmail.com (Log Out)
You've rated **225** movies.
You're the 46th visitor in the past hour.

★★★★★ = Must See
★★★★☆ = Will Enjoy
★★★☆☆ = It's OK
★★☆☆☆ = Fairly Bad
★☆☆☆☆ = Awful

Home | **Find Movies** | **Q&A (new)** | **Preferences** | **Help**

Shortcuts | **Search**

Rate and Find Movies

- **Top Picks For You**
- **Newest Additions**
- **Most Often Rated**
- **Rate Random Movies**
- **Browse Movies by Tags**

Your Movies

- **Your Ratings**
- **About Your Ratings**
- **Your Wishlist**
- **Your Tags**

Your Account

- **Your Profile (edit)**
- **Preferences**
- **Manage Buddies**
- **Manage RSS Feeds**

Help MovieLens

- **Volunteer Center**
- **Add a Title**

New Movies | **New DVDs**

★★★★½	Captain America: Civil War (2016)
★★★★	Ali Wong: Baby Cobra (2016)
★★★★	The Jungle Book (2016)
★★★★	The Man Who Knew Infinity (2016)
★★★★	Sing Street (2016)
★★★★	Weiner (2016)
★★★★	The Nice Guys (2016)
★★★★	The Crew (2016)
★★★★	Dreamland (2016)
★★★★	Veloce come il vento (2016)

200 new movies have been added since you last visited. See the newest additions.

Movie Tuner **Updated!**

Search Criteria: Similar to *Pulp Fiction*, but **more** action

Kill Bill: Vol. 1 (2003)	★★★★
Kill Bill: Vol. 2 (2004)	★★★★
True Romance (1993)	★★★★
Sin City (2005)	★★★★
Grindhouse (2007)	★★★½

★'s are MovieLens predictions. ★'s are your ratings.

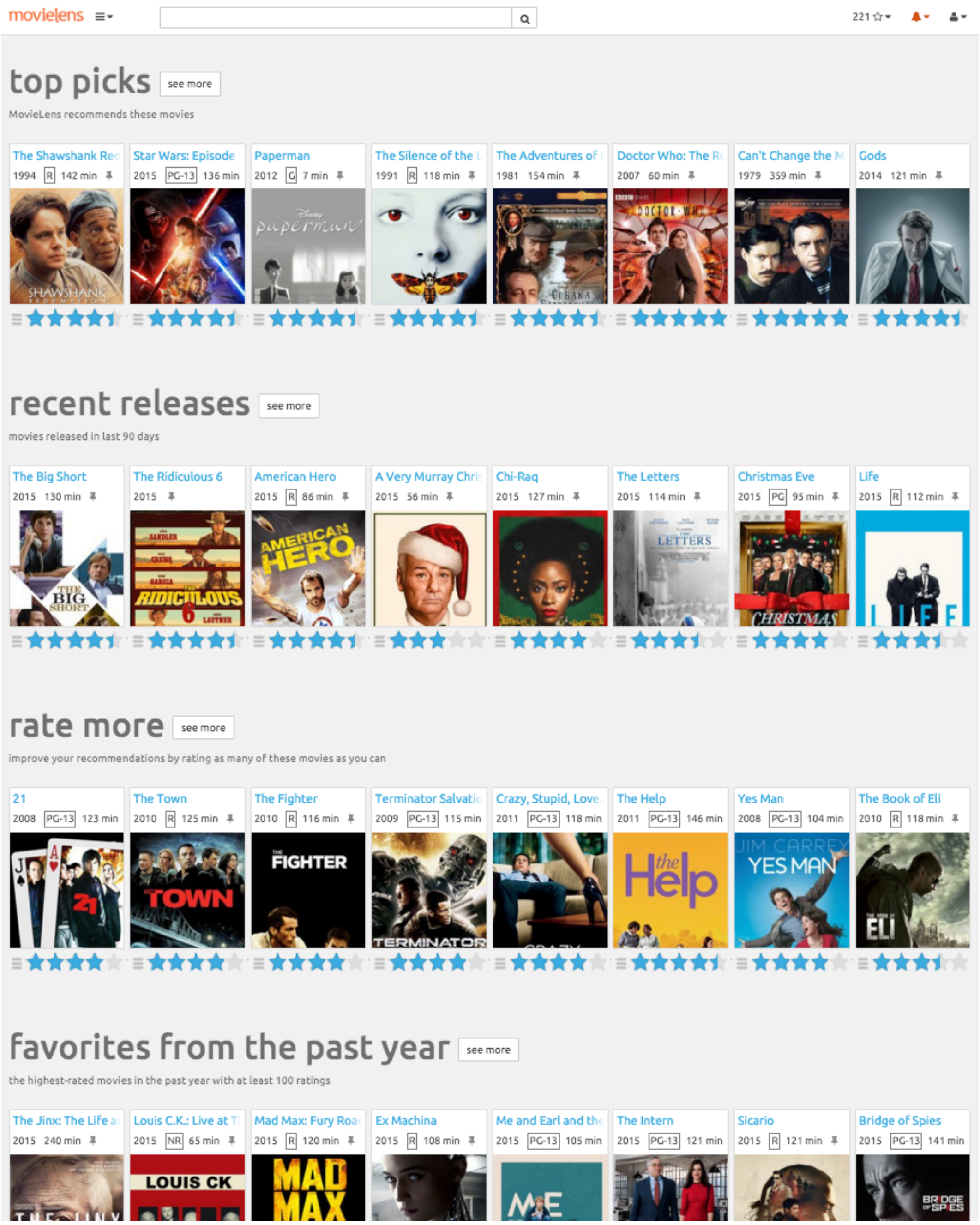
Want a movie like *Pulp Fiction* but less "violent"? Or a movie like *Mission: Impossible* but more "realistic"?

Movie Tuner lets you "tune" your movie selection along 1500 unique dimensions.

We recently updated Movie Tuner. We improved the list of default dimensions and added predictions to the results.

Look for Movie Tuner on the **Movie Details** and **Search Results** pages. More info

Figure 1.2: The homepage of the 3^{rd} version of MovieLens.



top picks

MovieLens recommends these movies

The Shawshank Rec 1994 R 142 min
Star Wars: Episode 2015 PG-13 136 min
Paperman 2012 G 7 min
The Silence of the L 1991 R 118 min
The Adventures of 1981 154 min
Doctor Who: The R 2007 60 min
Can't Change the M 1979 359 min
Gods 2014 121 min

recent releases

movies released in last 90 days

The Big Short 2015 130 min
The Ridiculous 6 2015
American Hero 2015 R 86 min
A Very Murray Chri 2015 56 min
Chi-Raq 2015 127 min
The Letters 2015 114 min
Christmas Eve 2015 PG 95 min
Life 2015 R 112 min

rate more

improve your recommendations by rating as many of these movies as you can

21 2008 PG-13 123 min
The Town 2010 R 125 min
The Fighter 2010 R 116 min
Terminator Salvatic 2009 PG-13 115 min
Crazy, Stupid, Love. 2011 PG-13 118 min
The Help 2011 PG-13 146 min
Yes Man 2008 PG-13 104 min
The Book of Eli 2010 R 118 min

favorites from the past year

the highest-rated movies in the past year with at least 100 ratings

The Jinx: The Life a 2015 240 min
Louis C.K.: Live at T 2015 NR 65 min
Mad Max: Fury Roa 2015 R 120 min
Ex Machina 2015 R 108 min
Me and Earl and the 2015 PG-13 105 min
The Intern 2015 PG-13 121 min
Sicario 2015 R 121 min
Bridge of Spies 2015 PG-13 141 min

Figure 1.3: The homepage of the 4^{th} version of MovieLens.

Chapter 2

Prior Work

In this dissertation, we focus on four aspects of user experience: the long-term effect of recommender systems on users, user preferences for diversity, popularity, and serendipity, the relationship between self-reported user satisfaction, the three recommendation properties with the inferred user recommendation consumption, and user experience with rating interfaces. We discuss the relevant related work in this section.

2.1 The long-term effect of recomender systems

While there is little debate about the efficacy of recommender systems in commerce, the debate about whether recommender systems are harmful to users has plenty of scholars on each side.

Pariser [19] argued that the root of human intelligence is the ability to adjust and adopt with new information and that recommender systems trap users into an unchanging environment. This unchanged environment, which he coined *the filter bubble*, reduces creativity and learning ability, and strengthens the belief of the user.

Tetlock [20], a political scientist, ran a study in which he asked different people with the various background for opinions on political and economic issues.

Surprisingly, he found that normal people gave more accurate predictions than the experts. A reason for the low prediction accuracy of the experts is that their views of the world are strengthened after years of study, leading to bias in making predictions.

Sunstein [21] went further and argued that by absorbing experiences that are personalized to them, users share fewer and fewer common experience with each other. He argued that '*Without shared experiences, a heterogeneous society will have a much more difficult time in addressing social problems. People may even find it hard to understand one another*' (p. 6, [21]).

On the other hand, Negroponte, co-founder of the MIT Media Lab, suggested that users can use recommender systems in such a way that it helps them to learn and explore new things. One such way is explored in 'The Daily US', in which users have personal intelligent agents that explore and summarize topics that are not the users' interests [22]. Negroponte called these intelligent agents *'the unequivocal future of computing'* [23].

Linden, one of the authors of Amazon's recommender system, suggested that narrowing user choices is not what personalization via recommender systems does. He argued that users can't search for items that they are not aware of, therefore, personalization increases serendipity [24]. With the idea of helping users achieve a good balance of awareness of new things, Kamba et al. [25] implemented a personalized news-agent called *Krakatoa Chronicle*. With *Krakatoa Chronicle*, users can choose how to balance between news that was personalized for them and news selected by editors as important for the whole community.

Fleder et al. [26], in their simulation study about the effect of recommender systems on sales' diversity, argued that at the user-user level, users are directed towards a common experience. This is because recommender systems cannot recommend items with little data (i.e. ratings), even if these items are favorable to the users. Therefore, recommended items can be new to an individual user, but they are overall the same (i.e. popular items). Hosanagar et al. [27], with a two-group designed study with users using iTunes as a recommender platform, also found that users tend to consume more common items. They argued that users consume the same items because recommender systems help users widen their interests, leading to higher chances of consuming the same items.

Although these studies present very interesting results regarding the debate about the filter bubble, they have some limitations. In Fleder et al.'s study, although they model user purchasing behaviors and how a recommender system works, their simulation does not capture the complexity of the user behaviors and their decision-making processes. Furthermore, with only two items in the simulation, their study cannot model the complexity of the eco-system of a recommender system, in which new items are added, and users' preferences drift over time. In Hosanagar et al.'s study, they build networks based on the user purchasing behaviors. They then compute the properties of these networks (e.g. median degrees & distances) as a measure if users tend to purchase the same songs.

Prior work leaves open the question of whether taking recommendations leads to narrowing of consumed content. To the best of our knowledge, our study is the first study looking at *recommended content diversity*, user behavior over the time, and the effect of taking recommendations on *consumed content diversity*. We describe our research in detail in chapter 3.

2.2 User satisfaction with recommendation diversity, popularity, and serendipity

In the previous section we discuss the importance of investigating the long-term effect of recommeder systems on users and the diversity of users' consumption. In this section we discuss what information of users recommender systems can use to potentially deliver individually preferred diversity levels, leading to an improvement in user satisfaction with recommender systems. Furthermore, we look at not only diversity, but also popularity, and serendipity because diversity, popularity, and serendipity are the fundamental properties of recommendations and recommendation lists. We start with a discussion about these three properties.

2.2.1 User preferences for recommendation diversity, popularity, and serendipity

Recognizing that maximizing accuracy in recommender systems may not lead to useful recommendations, researchers explore other metrics such as diversity, popularity, and serendipity.

Diversity addresses the problems of users receiving similar recommendations. These similar recommendations may not be useful to users. For example, users do not perceive a list of five recommendations to be useful when the list contains Toy Story I, Toy Story II, and Toy Story III. Popularity addresses the problems of users receiving popular recommendations. If the recommendations are popular, users may already know about them. Therefore, these recommendations are less useful. Serendipity addresses the problems of users not receiving unknown but interesting recommendations. Recommender systems should provide users with surprising but interesting recommendations.

Diversity indicates how different the recommendations are. It is measured as the average of all pairwise distances (representing the differences) of any two recommendations [28]. Serendipity indicates how unexpected but interesting the recommendations from what users usually consumed are. One way to measure recommendation serendipity is to first identify pairs of one item in the recommendation list and one item in the list a user has consumed. Then we take the average of all pairwise distances (representing the differences) of these pairs of items [29]. Recommendation popularity indicates how popular the recommendations are. One way to approximate recommendation popularity is to look at how frequent users consumed the recommended items [30].

We note that there are distinct differences between popularity and serendipity. A less popular item is not necessarily serendipitous for a user as he or she might not be interested in the item. On the other hand, a highly popular item is not necessarily less serendipitous for a user, because he or she might not hear about it yet due to some barriers (e.g. cultural barrier).

Prior work on diversifying recommendations argued that users would be more satisfied when they get diverse recommendations. Ali et al. [31] coined the term *portfolio effect* to describe the problem of recommender systems generating less diverse recommendation lists with only a few genres. Each recommendation in the set is good, but the set collectively is bad since it narrows on few genres. The incremental utility of recommendations for users decreases when users consume these similar items over and over. This phenomenon is termed *law of diminishing marginal returns* and is studied extensively in economics [32]. Indeed, Ziegler et al. [28] found that users liked diverse recommendations more than less diverse ones even though less diverse ones collectively have higher predicted ratings.

Similarly, prior work also suggested that recommending popular items is less useful to users since they probably already know about these items. Herlocker et al. [33] argued that popular recommendations were only useful when users were interested in these recommendations. In fact, recommender systems are found to be prone to popularity bias, presenting users with more popular recommendations than niche ones [30,34,35]. Users appreciate recommendation lists with some less popular items [36].

Likewise, prior work also suggested that recommendation lists should be serendipitous (i.e. be novel and interesting) to users to be useful to them. An example of useless recommendations is recommending bananas to grocery shoppers since everyone knows about bananas and perhaps always has intentions to buy them regardless [33]. Users will even be annoyed if they know about bananas but do not intend to buy them. Indeed, Zhang et al. [29] showed that increasing serendipity in recommendation lists improved user satisfaction.

2.2.2 User personality and user preference

We discuss above why the diversity, popularity, and serendipity metrics matter in recommender systems. We now look at the relations between user personality and user preferences for diversity, popularity, and serendipity of recommendations and recommendation lists.

Prior work in user personality have identified five personality traits - *openness to new experiences, conscientiousness, extraversion (or introversion), neuroticism and agreeableness [1,16,17].* Prior work also showed that there are significant connections between these traits and people's tastes and interests. Gosling et al. [18] argued that personality can be reliable sources to describe users' habits and behaviors. Kraaykamp et al. [37] found that personality has effects on users' media preferences. Rentfrow et al. [38] showed that personalities *"have roles to play in the formation and maintenance of music preferences".* More specifically, they showed that people with *openness to new experiences* usually tend to have preferences for jazz, blues, and classic music. Other studies showed that people who were *open to new experiences* preferred familiarity to novelty [17], or that *emotionally unstable people* people preferred popular media programs [37]. Chausson [39] also showed that people who are open to new experiences are likely to prefer comedy and fantasy movies while conscientious individuals are more inclined to enjoy action movies, and neurotic people tend to like romantic movies.

To the best of our knowledge, there has not been much research in the recommender system domain about the relationships between user personality and user satisfaction with the levels of recommendation diversity, popularity, and serendipity. Hu et al. [15] pointed out that there are few datasets with measures of user personality and their consumption (ratings). Among prior work on personality in recommender systems, Chen et al.'s work [40] is closest to ours. In their work, they reported that user personality influenced user needs for movie recommendation diversity. However, they focused on the diversities of genres, or actors, or countries of origin whereas our interest is the diversity of the movie contents as reported in [41]. Furthermore, they did not examine user satisfaction, user personality, and how user personality can improve user satisfaction. Therefore, in this study, we measure user personality and user evaluations of recommendation lists. We hypothesize that personality has significant connections with user preferences for the levels of diversity, popularity, and serendipity of recommendations, and how they enjoy those recommendations.

2.2.3 The assumptions of prior work in recommender system and our research opportunities

Although researchers in prior work demonstrated that user satisfaction highly correlates with recommendation diversity, popularity, and serendipity, they made two assumptions.

First, prior work assumed that users' ratings contain sufficient information about their preferences for diversity, popularity, or serendipity. Oh et al. [42] showed that recommender systems could learn about user tendency of consuming popular items via user ratings. They proposed a method called Personal Popularity Tendency Matching (PPTM) to measure these tendencies and used PPTM to generate useful recommendation lists. Zhang et al. [43] proposed a user-profile-partitioning technique to capture users' ranges of tastes for novelty from their ratings. They showed that this technique can improve recommendation novelty at a small cost to overall accuracy. Vargas et al. [44] also applied the same technique to improve recommendation diversity. We believe that users have their individual preferences for diversity, popularity, and serendipity and that we cannot depend on user ratings or user consumption data to infer these individual preferences.

Second, they assumed that users had global preferences for diversity, popularity, and serendipity. All users alike prefer diverse recommendations than unvaried ones, slightly less popular recommendations than highly popular ones, and more serendipitous recommendations than familiar ones. Thus, they focused on proposing methods to tune recommendations towards those global preferences. This assumption leads to contradicting results. For example, Zhang et al. [29] showed that increasing serendipity in recommendation improved user satisfaction, but Ekstrand et al. [45] showed that novelty reduced user satisfaction.

Thus, in chapter 4, we test these two assumptions. First, we investigate whether user preference for recommendation diversity, popularity, and serendipity can be extracted from user ratings alone. Second, we examine if users with different personalities have different preferences for recommendation diversity, popularity, and serendipity.

2.3 The relation between user satisfaction measured with survey questionnaires and users' consumption measured with logged data.

In the previous sections, we discuss the prior work on the long-term effect of recommender systems on users. We also present related work on what kind of information about users a recommender system can use to enhance user experience with regards to diversity, popularity, and serendipity of recommendations and recommendation lists. In this section, we discuss the relations between user satisfaction measured with survey questionnaires, and users' consumption measured with logged data.

To test out how good an approach to enhance user satisfaction is, researchers and recommender practitioners may carry out long-term experiments and compare user experience before and after the experiments. However, the associated risk of this approach is that users may be unhappy with recommendations in the experimental conditions and quit before the end of the experiment. Therefore, in chapter 5 we explore the relations between the self-reported user satisfaction with their consumption measured with longitudinal logged data. We also examine the relations between the three recommendation properties of recommendation sets measured with the measured consumption. The purpose of recommender systems is to suggest items that users subsequently enjoy consuming. Thus, understanding these relations helps us choose the appropriate measurement to infer users' future consumption without running the risk of losing users.

Most prior work in recommender system research employed survey questions to evaluate user satisfaction with recommendations. For example Ziegler et al. [28] used a survey to evaluate user satisfaction with diverse recommendations. Zhang et al. [29] asked users questions about the users' perceived enjoyment of serendipitous recommendations when they did a study about recommendation serendipity. Ekstrand et al. [45] also used survey questions to evaluate user satisfactions with recommendations generated by several different algorithms. Likewise, researchers use survey questions to assess other aspects of user satisfaction such as trust, effectiveness, or persuasiveness [46]. Survey Frameworks such as RecQue (Recommender systems' Quality of user experience) [47] or a Structural Equation Modeling based framework [48] are created for this purpose.

Although researchers can reliably measure user satisfaction at the time when the users evaluate the recommendations, minimizing the risk of imperfect recall on the users' responses [49], users often do not understand or are unable to explain their underlying motivations when answering survey questions. On the other hand, in the context of recommender systems where taking recommendations can be used to infer user satisfaction, we believe that understanding the relations between the self-reported user satisfaction and user recommendation consumption is valuable.

However, to the best of our knowledge, there has not been much research focusing on investigating and understanding these relations. Chen et al. [50] combined a survey questionnaire with a 3-week-period-controlled-field study in their research investigating the efficacy of four different recommendation algorithms. Our research interest, however, is in user satisfaction with recommendations. Our prior work [41] is one of the most recent research works establishing that post-recommendation consumption is a useful measure of recommendation list effectiveness. This recommendation list effectiveness implies user satisfaction with the recommendations. Another common approach is to measure click through rate (CTR), a ratio of clicked recommendations [51].

To the best of our knowledge, there is not much work bridging the self-reported user satisfaction with user recommendation consumption. Thus, we set out to look into these relations. We describe our study in chapter 5.

2.4 Improving user satisfaction via rating interfaces

We can also enhance user experience with recommender systems by improving their experience during the rating process. In this section we describe the related work about the rating process. We start with noise in user ratings.

Noise in user ratings is the undesired variances about the ratings. Such noise can occur because of inconsistencies in ratings over time [52]. There are two reasons for these inconsistencies. First, user preferences are unstable and therefore vulnerable to external influences such as anchoring effects [53] or manipulated predictions [54]. Second, users may have difficulties mapping their preferences into provided rating scales. Mapping their preferences accurately to the rating scales can be demanding for users, resulting in inconsistencies [55].

To address the noise and to increase the prediction quality, several researchers investigate the mapping process. From the algorithm perspective, reducing noise and improving predictions is possible using multiple ratings on the same user item-pair [56]. From the interface design perspective, changing to more granular rating-scales can help recommender systems collect and produce more information about user preference [57]. However, it has been shown that more granular rating scales demand more cognitive effort from users [58], and that gaining more accurate user models sometimes can lead to a decrease in user experience and user loyalty [59].

Prior work on decision making and preference construction and elicitation shows an *"emerging consensus"* [60] towards the idea that decision makers do not have well defined preferences [61, 62]. Preferences are not easily retrieved from some master list in one's memory, but rather are constructed while making a decision [63] and depend strongly on the situational context and task characteristics. Researchers have recently linked these phenomena to memory processes [64]. When a person is asked to evaluate an item, a memory query will be activated to recall information that is needed for the evaluation. This query will activate knowledge and associations of the particular item that will be used to form a preference of that item. In other words, the evaluation of the item will depend on how the preference for that item is (re)constructed from memory and from information provided in the interface.

Our work in chapter 6 explores and proposes mechanisms to reduce natural noise in user rating and facilitate users in their rating process, leading to improvements in the overall user experience.

Chapter 3

Effects of Recommenders on User Consumption Diversity[1]

In less than two decades, recommender systems have become ubiquitous on the Internet, providing users with personalized product and information offerings. They play a significant role in companies' profit margins. For example: Amazon once reported that 35% of its sales came from its recommendation systems [65]. Netflix in 2012 reported that 75% of what its users watched came from recommendations [66]. Recommender systems have greater influence on users' choices than peers and experts [67]. They lower users' decision effort, and improve users' decision quality [68].

But from the early days of recommender systems, researchers have wondered whether recommender systems might cause the *'global village'* to fracture into tribes [69], leading to *'balkanization'* [70]. Pariser [19] characterizes this worry in terms of a *'filter bubble'* – a self-reinforcing pattern of narrowing exposure that reduces user creativity, learning, and connection. We seek to examine the long-term effect of recommender systems on user consumption.

Investigating the filter bubble effect requires an access to a longitudinal dataset

[1] This chapter is based on our work [41] published in the 23rd international conference on World wide web.

that represents users' interactions with a recommender system and consumption of information items. We also must be able to distinguish users who act on the system's recommendations from those who do not.

In this paper, we meet these challenges by analyzing long-term users of MovieLens recommender system. We look at whether recommendations *received* become more narrow over time, but more important we also look at whether content consumed by using recommendation systems becomes more narrow. And because the essence of the risk of filter bubbles - that is people enter them willingly because they provide appealing content - we also explore the question of whether recommenders indeed provide that positive experience, leading their users to consume content they enjoy better.

3.1 Research Questions

We frame two specific research questions:

- **RQ1**: Do recommender systems expose users to narrower content over time?
- **RQ2**: How does the experience of users who take recommendations differ from that of users who do not regularly take recommendations?

To answer these questions, we develop two new research methods to isolate and measure the effect of accepting recommendations from recommender systems. First, we separate users into categories based on how often they actually consume recommended content. This separation lets us focus on users where a filter bubble is possible, and to compare these users against a control group who use the same system but do not regularly follow recommendations. Second, we introduce a method and metrics for exploring changes in the diversity of consumed items over time. This method looks at the items consumed (in our case, rated) in a time window, and then uses the *tag-genome* – a content coding derived from the community of users – to measure the diversity of those consumed items of a user.

Hence, these analyses let us address the question of the filter bubble where it is most relevant – at the individual level.

In answering these research questions, we make several contributions. First, we introduce a novel set of methods to study the effect that recommender systems have on users. Second, we provide a quantitative evidence suggesting that users who take recommendations receive a more positive experience than users who do not. Third, we find an evidence that while top-recommended items become more similar, the reduction in diversity is relatively small. Finally, we find that recommendation-takers consume more *content diverse* movies than non-recommendation-takers, and that these users are actively seeking to watch more diverse movies.

3.2 Data and Metrics

In this section we describe our datasets and discuss our methods for identifying recommendation takers and computing the content diversity of movies.

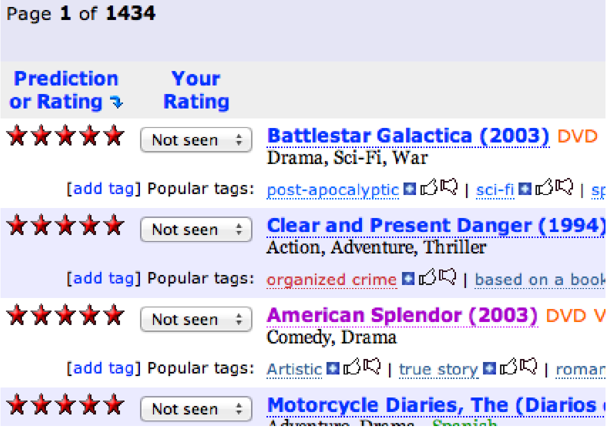
3.2.1 Dataset

To answer our research questions, we use data from MovieLens.[2] MovieLens is a movie recommender system that has been in continuous use since 1997. As of September 2013 , there are 217,267 unique users who have provided more than 20 million movie ratings for more than 20,000 movies. We use this data because it offers us three unique advantages: longitudinal data, a recommender system with a well-known recommender engine, and an expressive way to compute content diversity.

Longitudinal data. MovieLens provides us longitudinal data of user rating data. MovieLens logs capture timestamps and other information when users rate

[2] The data can be downloaded from http://www.movielens.org/. We used data from the third version of MovieLens as mentioned in 1.3.

movies and when they view pages of recommended movies.



Page **1** of **1434**

Prediction or Rating | **Your Rating**

★★★★★ Not seen **Battlestar Galactica (2003)** DVD
Drama, Sci-Fi, War
[add tag] Popular tags: post-apocalyptic | sci-fi | sp

★★★★★ Not seen **Clear and Present Danger (1994)**
Action, Adventure, Thriller
[add tag] Popular tags: organized crime | based on a book

★★★★★ Not seen **American Splendor (2003)** DVD V
Comedy, Drama
[add tag] Popular tags: Artistic | true story | romar

★★★★★ Not seen **Motorcycle Diaries, The (Diarios**
Adventure, Drama, Spanish

Figure 3.1: Top Picks For You

MovieLens provides a feature called *'Top Picks For You'* (shown in figure 3.1) that takes users to a page displaying movies the users have not seen, ordered from the highest predicted ratings to the lowest predicted ratings. By default, a *'Top Picks For You'* page displays 15 movies, though users can change this default number[3]. Since May 2003, MovieLens started to log all user access to *'Top Picks For You'* pages and recommended movies with their respective positions in the recommendation lists at the time users accessing the page. Knowing when and what movies users rated, and when and what was recommended to users helps us identify if users are taking our recommendations and how consistently they consume the recommendations.

A recommender system with a well-known recommender engine. MovieLens uses an item-item collaborative filtering (CF) algorithm[4], a well-known and broadly-used recommendation algorithm that is robust in performance and scalability with high dimension data [71]. Due to these advantages, Amazon - one of the early industrial recommender systems - used it in production [72]. We think that analyzing the longitudinal data generated from one of the well-known

[3] Our analyses suggest that only 3.2% of our users change this number.

[4] MovieLens switched to item-item CF algorithm in 2003.

and broadly-used algorithms makes our case more generalizable.

An expressive way to compute content diversity. MovieLens provides tag-genome data, an expressive way to characterize movie content. *'Tag-genome'* is an information space in which for any pair of a movie and a tag, a relevance score is computed to indicate how best the tag describes the movie. Since 2006, MovieLens has provided a feature that allows users to apply tags (words or short phrases) to movies. Vig et al. [14], based on this tagging feature and the tags that MovieLens users have applied, built tag-genome to help users navigate and choose movies where all dimensions, but one, are the same as those of the compared movie. In section 3.3, we will describe the tag-genome data in details and illustrate why we use this data to measure content diversity.

At the time we took a snapshot of the tag-genome data (April 2013), it consisted of 9,543 distinct movies described by 1,128 distinct tags (10,764,504 pairs). In our analyses, all of the movies are in this information space.

In this study, we analyze data in the period from February 2008 to August 2010 (21 months).

3.2.2 Identifying recommendation takers

To study the effect of 'taking' recommendations, we need to classify users in our dataset into those that do 'take' recommendations and those that do not. In this section, we describe how we define these two groups.

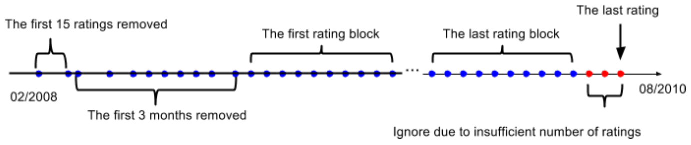

The first 15 ratings removed
The first rating block
The last rating block
The last rating
02/2008
08/2010
The first 3 months removed
Ignore due to insufficient number of ratings

Figure 3.2: Rating Block Ilustration

Rating Block. Our objective in this study is to examine the temporal effect of recommender systems on users throughout their lifecycles. To do so, we have to divide the rating history of a user into discrete intervals.

Before we define these smaller intervals, for each user we remove the first 15 ratings because these ratings are given based on the movies suggested by MovieLens in order to gain knowledge about the preference of the new user. Then, we remove all of the ratings from the first three months after the first 15 ratings. We do this for three reasons:

- some users rated an abnormally high numbers of movies in the first three months. This is potentially due to the fact that these users had watched many movies before joining MovieLens – they rated these movies to help MovieLens understand their preference better. However, in this study, we want to capture the consumed movies which were recommended;
- we want to give users sufficient time to learn how to use MovieLens;
- we want to give MovieLens enough time to understand users' preferences better, in order to improve the quality of recommendations.

After removing these initial movies, we formulate intervals for the remaining rating history of a user. There are several ways to define an interval. One is to define an interval as a login session. Another is to define an interval as a block of n consecutive months. However, both of these approaches possess some potential problems for our temporal analyses. First, the numbers of ratings per interval among users are different. These differences are because the frequencies in using recommender systems are different among users. Second, the numbers of ratings provided by users diminishes over time. Hence, with an interval defined as n consecutive months (or even as n logging-in sessions), some intervals will have a lot of ratings, at others will not. These problems potentially make our analyses unreliable because the effect of recommender systems are different in different intervals.

To address both of the potential problems described above, we define an interval as a block consisting of 10 consecutive ratings[5]. With this definition, from

[5] We also analyzed with other block sizes (e.g. a block consisting of 15 (or 5) consecutive ratings),

now on we refer to an interval of 10 ratings as a *rating block*. With this constant number of ratings per block, we make sure that all users rated the same numbers of recommendations throughout a defined rating block. We choose 10 ratings per rating block because we want a rating block sufficiently long enough to capture the long-term effect, and because our analyses show that 10 is the median of the distribution of numbers of ratings per 3 months, a sufficiently long time interval. If there are not enough ratings to form the last rating block, we will drop these ratings because we want to make sure all rating blocks have the same number of ratings. Figure 3.2 summarizes our method of forming a rating block.

We only select users whose first ratings were in the *analyzed period* (i.e. in the period of February 2008 - August 2010 as defined in the section 3.1). We include only those users who have three or more ratings blocks in the analyzed period. To simplify our writing, in the remainder of this paper we refer to this selected group of experimental subjects simply as users.

Overall, we have 1,405 users in our analyses. These users made at least 3 rating blocks and at most 203 rating blocks (mean$= 12$, $\sigma = 15$). In our analyzed period, February 2008 to August 2010, the 1,405 users provided 173,010 ratings on 10,560 distinct movies. 100% of these movies are in tag-genome database described above. They accessed their *'Top Picks For You'* 150,759 times.

Identifying consumed recommendations in a rating block. To investigate the effect of recommender systems on users, we need to identify which movies in each rating block were explicitly recommended to the user in the interface. With these recommended movies identified, we can measure the level of recommendation intake of a user during his rating history. Furthermore, with recommended movies identified in a rating block, we can examine the user experience when taking and not taking recommendations at the same time (i.e. within a rating block). Based on individual levels of recommendation intake, we classify users into two groups - those who take recommendations and those who do not. We will discuss our

and we observed the similar results.

classification method in more detail in the next section.

We define if a movie was recommended to user u by checking if the movie was displayed in the *'Top Picks For You'* before. Specifically, for any user u, a movie in his i^{th} rating block is defined as *recommended to him* if and only if the movie was in *'Top Picks For You'* between 3 hours and 3 months before user u rated this movie. Figure 3.3 visualizes our definition.

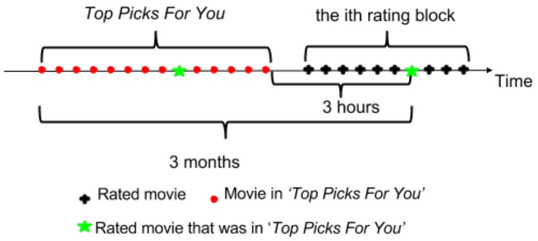

Top Picks For You
the ith rating block
Time
3 hours
3 months
✚ Rated movie
● Movie in 'Top Picks For You'
★ Rated movie that was in 'Top Picks For You'

Figure 3.3: Identifying if a rated movie was recommended before.

We require at least three hours to avoid the case where user u rated a movie upon seeing it in his *'Top Picks For You'* (an indication that the user rated it because they had seen it previously, not because they took the recommendation and watched it on the spot). We believe that three hours is sufficient time for a user to watch a movie, then rate it. We set a limit of three months to accomodate the fact that some users might need substantial time to rent and consume a movie; we capped the time limit to accomodate the reality that as time passes, the likelihood of a causal link between the recommendation and consumption diminishes.

In the next section, we discuss how we classify our users into two groups - a group that took recommendations (*Following Group*) and a group that did not (*Ignoring Group*).

Ignoring Group v.s. Following Group The purpose of our study is to investigate the long term effect of using recommender systems on content diversity. To this end, it is useful to draw comparisons between two groups of users - one that consumes recommendations consistently over time, and one that does not.

Suppose that we classify user u solely based on the ratio of his rated movies that were recommended over the number of the rated movies in his rating history. Some users might always take recommendations towards the beginning of their rating histories, then do not take any recommendations towards the end. With potentially high ratios, these users could be classified as recommendation takers. However, the effects of the recommender systems on these users are only towards the beginning of their rating histories.

In order to estimate the consistent recommendation intake of a user over his rating history, we first look at whether the user took at least one recommendation in one of his rating blocks using the proposed method in the previous section. We argue that as long as within a rating block, user u took a recommendation, there was an effect of the recommendation system on that user in that rating block. We then compute the percentage of that user's rating blocks in which the user took at least one recommendation.

With these per-user percentages computed, we rank our users from the highest percentage to the lowest percentage. That said, the users who took recommendations in all of their rating blocks (i.e. percentage = 100%), are placed on top, those that did not take recommendations in any of their rating blocks (i.e. percentage = 0%) are placed bottom. Users who did not take any recommendations in any of their rating blocks are classified as non-recommender takers and placed in the *Ignoring Group.* Users who took recommendations in at least 50% of their rating blocks are classified as recommender takers and placed in the *Following Group.* Overall, the *Following Group* consists of 286 users, and the *Ignoring Group* consists of 430 users. Of these 430 users in the *Ignoring Group*, 52 never access to *'Top Picks For You'*[6] and 378 accessed *'Top Picks For You'* but never consumed any of these recommendations. Figure 3.4 visualizes our classification method.

[6] Harper et al. [73] showed that not all users came to a recommender system for recommendations.

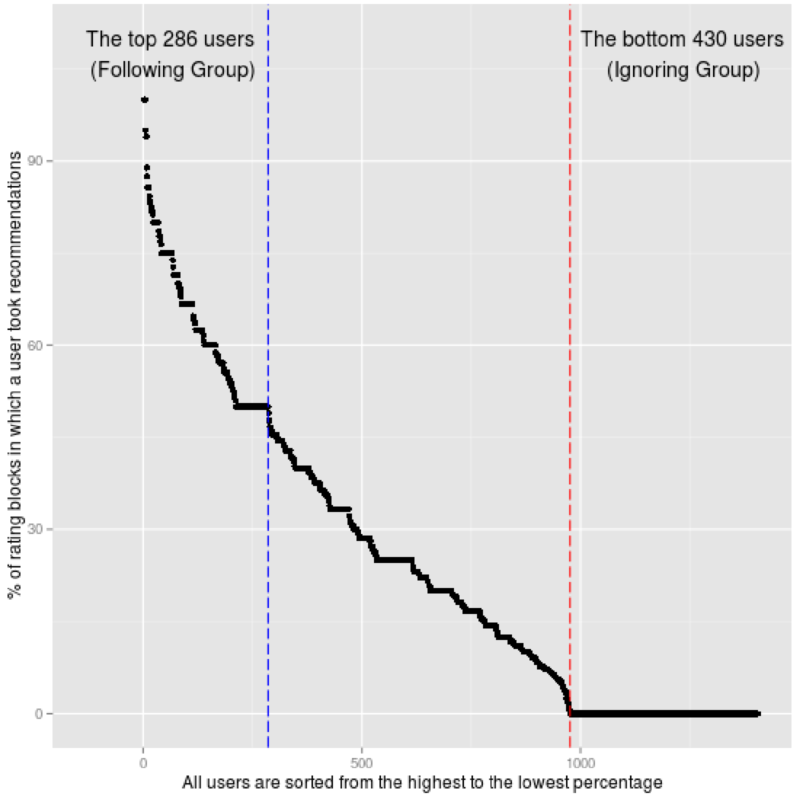


Figure 3.4: The visualization of our methodology to identify recommendation takers and non-recommendation takers. All users are sorted from the highest to the lowest percentages. The two cut-off points blue and red are at 50% and 0% respectively.

3.2.3 Measuring Content Diversity

Our study examines the effect of recommender systems on the content diversity of recommended and consumed (rated) movies. In this section, we describe the tag-genome data, our method to compute content diversity using tag-genome, and discuss why we use tag-genome.

The *tag-genome* [14] is an information space containing a set M of movies, and a set T of tags. The set T consists of tags that are highly descriptive about movies. How well tag t describes movie m is expressed via the relevance score

$rel(t, m)$. The relevance score $rel(t, m)$ takes on values from 1 (does not describe the movie m at all) to 5 (strongly describes the movie m)[7]. Each movie m_i is represented as a vector of size $|T|$ where entry i, j is the relevance of tag j to movie i. Please refer to 1.3.2 for more details.

To measure the similarity of two movies, we compute the Euclidean distance between two movie vectors. That is:

$$d_{(m_i,m_j)} = \sqrt{\sum_{k=1}^{m}[rel(t_k, m_i) - rel(t_k, m_j)]^2}$$

Lower numbers indicate greater similarity. We use Euclidean distance instead of cosine distance because the movie × tag-genome matrix is dense (i.e. $rel(t_k, m_i) > 0 \quad \forall i, k$). The minimum distance in MovieLens dataset[8] is 5.1, representing the distance between two movies in the *'Halloween Series'* : *'Halloween 4: The Return of Michael Myers (1988)'*, and *'Halloween 5: The Revenge of Michael Myers (1989)'*. The maximum distance in MovieLens dataset is 44.24, representing the distance between two movies *'Paris was a woman'* and *'The Matrix'*. The average distance is 23.44 representing the distance between two movies *'Chronicle (2012)'* and *'End of Watch (2012)'*. The standard deviation of movie distances is 4.45.

We use tag-genome because it provides an expressive way to describe the content of a movie. This expressive way is better than the traditional method of computing movie content diversity via user rating vectors. If two movies have similar user rating vectors, that means they are similarly liked, not that their content is similar. It is also better than computing movie content diversity based on meta-data such as genres, actors, or directors, etc., because two movies that share actors or directors (or even are *' comedies'*) may not actually be similar.

The strength of our tag-genome-based method lies in how tag-genome computes the relevances between the set of tags T and the set of movies M. These

[7] Vig et al. [14] originally proposed the range to be from 0 to 1. However, after several revisions, as of 09/26/2013 MovieLens uses the range of 1 to 5.

[8] The data from the third version of MovieLens.

relevances are computed based on a community-supervised learning approach. In this approach, users provide the training dataset by evaluating how strongly a tag describes a movie. With the training dataset and other sources of tags such as IMDB, MovieLens predicts the relevances for other pairs based on different machine learning models. Furthermore, the relevance of any pair of tag t and movie m is constantly refined via feedback from users. Hence, these relevance scores are better at describing the content of movies than user rating vectors and properties such as genres and directors. Due to its unique advantages, researchers have shown that the *tag-genome* can help users navigate through a collection of thousands of movies [74], and can assist users in remembering what movies are about [75].

Tag	*‘Halloween 4 ...’*	*‘Halloween 5 ...’*	*‘The Front P...’*
creepy	2.551	2.328	1.087
revenge	3.185	3.920	1.311
franchise	4.992	4.993	1.162
suspense	3.864	3.890	1.261
nudity (topless)	2.712	2.848	1.071
supernatural	3.386	3.434	1.055
serial killer	4.940	4.943	1.065
splatter	2.089	3.439	1.435
teen movie	3.450	3.949	1.347

Table 3.1: The 9 tags describe the three movies.

To illustrate the difference of using user rating vectors and the *tag-genome* for computing content diversity, we look at the following example. Based on the tag-genome, the movie that is the most content similar to movie *‘Halloween 4: The Revenge of Michael Myers’* is *‘Halloween 5: The Revenge of Michael Myers’*. However, based on the user rating vectors of MovieLens data, the most similar movie to *‘Halloween 4 ...’* is *‘The Front Page’* (with the cosine similarity is 0.991). Clearly, taking the content similarity in consideration, it is obvious that the former is more accurate than the later because *‘Halloween 5 ...’* is the fifth movie in the ‘Halloween film series’ whereas *‘The Front Page’* has a different story line. Table 3.1 shows 9 tags that described two movies *‘Halloween 4 ...’* and *‘Halloween 5*

...') but not *'The Front Page'*.

3.2.4 Measuring The Effect of Recommender Systems

In this study, we measure the effect of recommendation systems on content diversity as well as the user experience. In this section, we describe the metrics to compute content diversity and user experience. Then, we discuss how we measure the effect of recommender systems.

The Metrics

Content Diversity. We compute the content diversity distribution of a group of users by computing the *movie distance* distribution of the group. Specifically, the content diversity of a list of recommended movies to user u is the average pair-wise distances of the movies in the list. We also do the same to compute the content diversity of rated movies. Measuring the diversity of a list of items by averaging pairwise diversity scores was developed by Ziegler et al. [28]. To make our study more robust, we also use the maximum value of the pair-wise distances as the content diversity metric. In our results, we will report both the average as well as the maximum pair-wise distances of a list of movies.

For recommended movies, we compute the content diversity of the top 15 recommended movies per user. We choose only the top 15 because for most of our users who consulted *'Top Picks For You'*, MovieLens always captured at least the top 15 recommended movies for them. This is because 15 was the default number of recommended movies shown on the first page when a user clicked on *'Top Picks For You'*, Furthermore, only 0.05% of the MovieLens' users changed the default number to less than 15.

For the consumed movies, we measure the content diversity of rated movies. Since we divide a user history into smaller rating blocks, we compute the content diversity of all 10 rated movies in a rating block for all rating blocks.

User Experience. For user experience, we measure how much users enjoy movies via their given ratings in MovieLens. Specifically, we compute per user rating average of movies in a given rating block.

Group Comparison

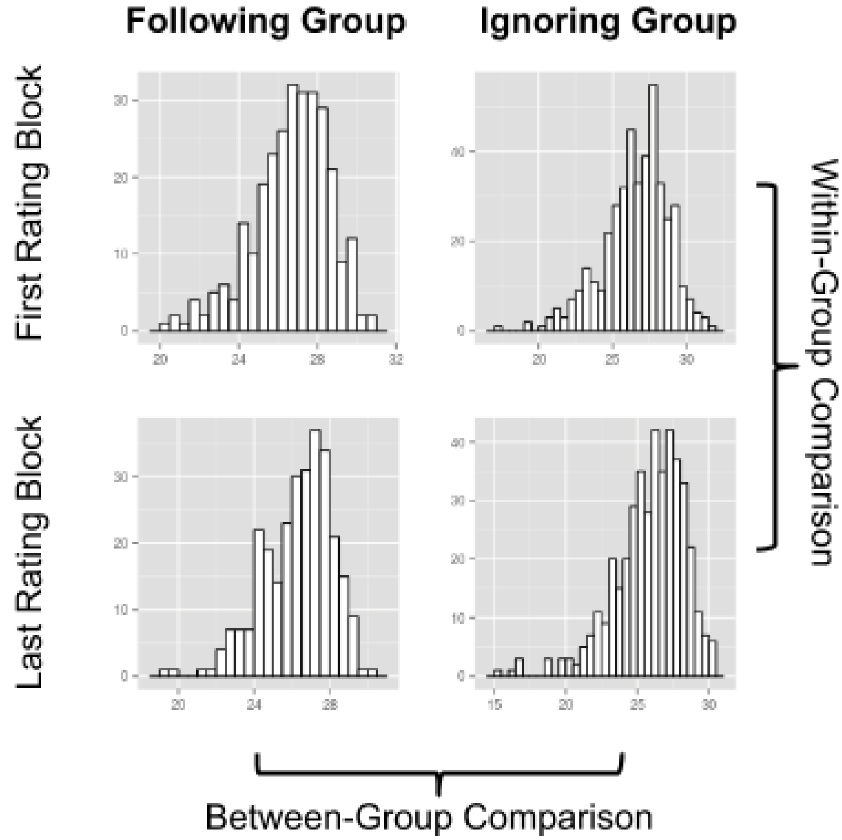


Figure 3.5: The visualization for our within and between group comparisons for the content diversity of the consumed movies, where x is movie distance and y is a number of users.

Since all the content diversity and user experience distributions are approximately normal, we investigate the effect of recommender systems by measuring the shift in means of the two distributions of a group.

Specifically, to examine the effect of recommender systems on content diversity of a group, we measure the shift in means of the content diversity distributions at the beginning and at the end of the rating histories of all users in the group. We do the same for measuring the effect on user experience. Comparing the effects of recommender systems at the beginning and the end of a rating period is used by other researchers (for example Hosanagar et al. [27]). We call this within-group comparison.

To examine how the effect on content diversity or user experience is different between two groups, we measure the shift in means of the distributions of the two groups at the beginning rating histories of all users in the group. We do the same for the two distributions of the group at end of user rating histories. We call this between-group comparison. Figure 3.5 visualizes our comparison method.

Since all the content diversity (i.e. movie distance) distributions are approximately normal, we use t-tests to compare the means of the two distributions. Specifically, for the within-group comparison, we use a paired t-test. For the between-group comparison, due to the different sizes of the populations, (286 users in the *Following Group* vs. 430 users in the *Ignoring Group*, we use Welch's t-test. All the above t-tests can be performed using the R statistical package[9]. We do the same for the user experience distributions.

3.3 Results

We present our results as they relate to our two research questions.

RQ1: Do recommender systems expose users to narrower content over time?

To answer this question, we compare the content diversity of recommended movies at the beginning and at the end of a user's observed rating history.

Table 3.2 shows the content diversity of all users[10]. We observe that for all

[9] http://www.r-project.org

[10] Of 1405, 4 changed the default number to less than 15; 52 users never accessed to the *'Top Picks*

users the average pair-wise distance of the top-15 recommended movies becomes smaller over time with a drop from 25.02 to 24.67. The p-value for the t-test is 2.43e-06, showing that the difference in the means between the two distributions are statistically significant. Therefore, although the drop in content diversity of the recommendations is small, it is statistically significant.

	At the beginning	At the end	Within-group p-value
All users	25.02	24.67	2.43e-06
Following Group	25.22	24.80	0.014
Ignoring Group	24.74	24.51	0.087
Between-group p-value	0.0037	0.0406	

Table 3.2: The average content diversity of the top 15 recommended movies

These drops in content diversity are also observed in *Following Group* as well as the I͡gnoring Group with the within-group p-values are 0.014 and 0.08 respectively. That means for the movies recommended to the *Following Group* became more and more similar ($p < 0.05$). Although for the *Ignoring Group* the movies recommended became more and more similar, the trend was marginally significant. The differences for the two groups have a significant meaning. The *Ignoring Group* did not take recommendations from *‘Top Picks For You’*, leading to minimal changes in the recommendation lists. This change is due to the fact that MovieLens still learned about the preferences of these users via ratings. MovieLens then made adjustments in the recommendation lists, and recommended movies that were more similar to these users’ preferred ones.

Interestingly, we also observe that the recommended movies to the *Following Group* seems to be more content diverse that those recommended to the *Ignoring Group* (the between-group p-value = 0.0037 at the beginning and p-value = 0.0406 at the end of user rating histories). However, the difference in the content diversity of the two groups becomes smaller over time (0.48 at the beginning v.s. 0.29 at the end). Eventually the content diversity of the *Following Group* may become less than that of the *Ignoring Group*. However, this is an issue for future work.

For You‘. Thus the number of users analyzed for this analysis is 1349.

Negroponte [22], Linden [24], Kamba et al. [25], and other researchers have proposed that users can use recommender systems as tools to explore new things that they are not aware of. Hence, potentially the content of consumed movies might be diverse. Thus, it is of our interest to investigate how taking recommendations affects the users' consumed content diversity and user experience. In the next section, we set out to answer our second research question:

RQ2: How does the experience of users who take recommendations differ from that of users who do not regularly take recommendations?

To answer our second research question, we set out to answer the following questions:

a) Does taking recommendations lower the consumed content diversity?

Our results, as shown in table 3.3, suggest that at the beginning, there is no difference in the content diversity of the consumed (rated) movies by the two groups (26.67 vs. 26.59 with p value of the t-test = 0.6162). This suggests that, after using recommender systems for the first three months[11],the effect of recommender systems on the consumed movies of both groups is not significantly different.

Rating Block	The First	The Last	Within-group p-value
All users	26.60	26.01	1.542e-12
Following Group	26.67	26.30	0.01007
Ignoring Group	26.59	25.86	8.236e-07
Between-group p-value	0.6162	0.006468	

Table 3.3: The average content diversity of the consumed movies of the two groups

However, our results also suggest that after using MovieLens for about nine months[12], we can see the effect on content consumed by users. At the end of our

[11] We recall that the first three months of usage history are removed before forming the first rating block (see section 3.2).

[12] The initial first 3 months for users to learn and efficiently use the system, also for a recommender

observed periods, the content diversity of both groups is reduced. With p-values of approximately zero showing that the reductions are significant. Interestingly, we also observe that compared to *Following Group*, the *Ignoring Group* had higher drop.

We observe similar results when we define the content diversity as the maximum distance of a pair movies in the movie list (table 3.4). Using this metric, we find no differences between the two groups during the first three months (p-value = 0.237), and we find that users consume less diverse movies over time (p-value = 8.903e-07). Again, the following group consumed more diverse content than the ignoring group.

Rating Block	The First	The Last	Within-group p-value
All users	34.56	34.00	8.903e-07
Following Group	34.73	34.36	0.127
Ignoring Group	34.45	33.73	0.000
Between-group p-value	0.237	0.008	

Table 3.4: The maximum content diversity of the consumed movies of the two groups

Given the finding that the *Following Group* watched more diverse movies than the *Ignoring Group*, we ask:

	Rating Block	0.5 - 1 stars	1.5 - 2 stars	2.5 - 3 stars	3.5 - 4 stars	4.5 - 5 stars
All Users	The First	2.7%	5.3%	17.8%	46.5%	27.7%
	The Last	2.8%	6.3%	22%	46.4%	22.5%
Following Group	The First	2.2%	6.0%	17.8%	46.2%	27.8%
	The Last	1.8%	5.1%	19.0%	49.2%	24.9%
Ignoring Group	The First	2.4%	4.6%	18.0%	45.3%	29.7 %
	The Last	3.6%	6.9%	21.5%	45.1%	22.9%

Table 3.5: The percentage of rated movies in the respective rating ranges.

b) Did the ***Following Group*** have better experience?

By their nature, movies recommender systems help users find movies that they may enjoy. Enjoyment is expressed via ratings: the higher the rating, the more

system to pass the cold start phase, then the next 3 months for rating 10 movies (one rating-block) and the next 3 months for rating other 10 movies (another rating-block).

enjoyable the movie. However, we observe that for all users (N = 1,405), the rating averages at the first rating block and at the last rating block are 3.69 and 3.57 respectively, suggesting the drop of 0.12. This drop surprised us since we expected that recommender systems should have helped users identify movies better suited to their tastes.

To analyze the user experience further, we look at the percentage of movies all users rated at the rating scale from 0.5 to 5 stars as shown in table 3.5. The percentages of watched movies that were rated higher or equal to 3.5 stars drop (from 74.2% to 68.9%), whereas for the other rating stars, the percentages increase. We observe that overall, our users watched less enjoyable movies.

Interestingly, we observe that the *Following Group* consumed more enjoyable movies. The percentage of the movies rated from 3.5 - 4 stars for this group increases from 46.2% to 49.2%, and that of the movies rated from 0.5 - 1 star decrease from 2.2% to 1.8%. Furthermore, the percentages of the movies rated from 4.5 - 5 stars of *All Users* and *Ignoring Group* receives higher drop than that of the *Following Group* (5.2%, 6.8% and 2.9% respectively).

Group	Rating Block	In Predictions	0.5 - 1 stars	1.5 - 2 stars	2.5 - 3 stars	3.5 - 4 stars	4.5 - 5 stars
Following Group	The First	Yes	1.6%	3.8%	9.6%	50.2%	34.8%
		No	2.5%	6.1%	19.4%	45.4%	26.6%
	The Last	Yes	1.6%	2.9%	10.8%	52.0%	32.7%
		No	2.2%	5.2%	19.8%	48.8%	24.0%

Table 3.6: The percentage of rated movies in the respective rating ranges.

To verify that the trend - the users in the *Following Group* watched more enjoyable movies than the users in the *Ignoring Group* - is statistically significant, for each group, we compute the distributions of per user's rating mean in the first rating block. We also do the same for ratings in the last rating block.

Since these distributions are normal, and have the same number of users with approximately the same variance, we perform t-test on these distributions. Like the methodology visualized in figure 3.5, we compare the between-group distributions, and within-group distributions. However, this time our distributions are the distributions of rating mean.

Rating Block	The First	The last	Within-group p-value
All users	3.69	3.57	2.2e-16
Following Group	3.69	3.68	0.7
Ignoring Group	3.74	3.55	3.128e-11
Between-group p-value	0.2129	0.001719	

Table 3.7: Rating Mean of the two groups

Our results, as shown in table 3.7, suggest that in the first rating block, the users in the *Ignoring Group* had better experience than those in the *Following Group*. However, the enjoyment difference between the two groups (measured by the difference in rating mean) is not statistically significant at the 95% level of confidence interval (p-value = 0.2129). However, in the last rating block, the *Ignoring Group* watched less enjoyable movies than the *Following Group*. The enjoyment difference between the two groups is statistically significant (p-value = 0.001719). Furthermore, although the ratings of the *Following Group* decreases over time, this drop is not statistically significant (p-value = 0.7). For the *Ignoring Group*, the drop in ratings is statistically significant (p-value = 3.128e-11).

We look further into the experience users receive when they consumed movies that were and were not recommended. Specifically, we look at the experience of the *Following Group* since this group consumed a significant amount of recommended movies. As shown in table 3.6, we observe the users in the *Following Group* consumed more enjoyable movies. The group gave at least 3.5 stars for 85% and 84.7% of consumed recommended movies in the first and the last rating block respectively. On the other hand, the group only gave 72% and 72.8% of consumed non-recommended movies at least 3.5 stars in the first and the last rating block respectively. These numbers mean that the group received worse experience when watching movies that were not recommended for them.

These results suggest that the users who followed recommendations received a better experience than those who did not follow the recommendations. However, as we mentioned above, some users in the *Following Group* did not always take recommendations. To verify whether taking recommendations indeed improves the experience of a user, we seek to answer the following question:

c) What does the change of rating average mean?

To clearly understand what it means when the rating changes 0.12, we define the positive experience index as the percentile of per use rating average. That means the change in percentile of a user (or a group) indicates how the positive experience of that user (or a group) changes is comparing to the population. If the average rating of a user is at 90th percentile, that means he receives more positive experience than the rest of 90% of the population.

	Rating mean change	Percentile change
All users	-0.12	-11.97
Following Group	-0.01	-1.20
Ignoring Group	-0.19	-18.86

Table 3.8: The change in percentile corresponding the change in rating mean.

With this analogy, we build the percentile table based on the ratings of all of the users (N = 1,405) in the analyzed period (i.e. from February 2008 to August 2010). We observe that overall at the first rating block with the average rating of 3.69, the *Following Group* is at 58.93^{th} percentile. That means the group had a better experience than more than half of the population. Whereas, the *Ignoring Group* in the first rating block with the average rating of 3.74 (at 63.63^{th}*percentile*), had even better experience than the *Following Group*. However, in the last rating block, the percentile of the *Ignoring Group* drops to 44.77^{th} percentile, a 18.86 drop whereas the drop of the *Following Group* is 1.21. That implies as time went by, the *Ignoring Group* received worst experience, or watched significantly less enjoyable movies, than the *Following Group* did. Table 3.8 summarizes our results.

3.4 Discussion

We set out to better understand the broadening or narrowing influence of an online recommender system on its users: did it tend toward a filter bubble? We found evidence for two forms of narrowing when analyzing all users - the items recommended by the system and the items rated by users both became slightly narrower (less diverse) over time. However, the results for *all* users obscure the most interesting part of the story. The narrowing effect actually was mitigated for users who appeared to "follow" the recommender (operationalized as having rated movies that appear in their top-n recommendation lists); in other words, *taking recommendations lessened the risk of a filter bubble.*

First, recommendation-following users received more diverse top-n recommendation lists than non-following users. Because recommenders are personalized, user actions affect their output. In the case of the relatively standard item-item algorithm evaluated in this research, rating recommended movies (rather than movies chosen via other means) appears to encourage the algorithm to broaden its future recommendations. Second, recommendation-following users narrowed the content diversity of their rated movies more slowly – these users were still narrowing (significantly, but slightly), but the effect was smaller than for those users who never rated the movies the recommender showed to them.

This begs the question - is there a "natural" narrowing effect over time, at least in the domain of movies? After all, we form habits based on what we've watched recently, and as we watch more, we solidify our preferences. In the movie domain, we face the additional possibility that the best movies are relatively diverse in content, but limited in number; once we get through those, we turn to newer movies closer to our comfort zone. If this is true – if there is a natural tendency to narrow our consumption of movies (or other media) over time – then collaborative filtering-based recommenders appear to help mitigate the tendency, and thus may play a broadening role.

What can the designers of recommender systems do to discourage the narrowing tendency? First, they can use collaborative filtering algorithms like those in MovieLens, which slows the narrowing effect over time. It is an open question if content-based recommenders have the same effect as collaborative recommenders; we suspect the content-based alorithms will more strongly push users towards narrow consumption. Second, recommender systems can inform users about the diversity of their consumption. Be it movies or news, a site can display diversity metrics or summary statistics that help users better understand if they have in fact gone too far into a particular interest of theirs. Finally, if recommenders aren't enough to reduce the narrowing effect, we should explore further steps to intentionally increase diversification of recommendation lists. This is consistent with Ziegler et al.'s finding [28] that diversification can improve user satisfaction.

Our work has several limitations. We cannot be sure if people are really following MovieLens recommendations, since we are using log data analysis methods. Additionally, users may be influenced by recommendations from other information sources or from their friends. To verify recommendation-following behavior would require contacting users to develop baseline measures for recommendation awareness. Furthermore, due to the design of the MovieLens logging infrastructure, we are restricted to analyzing the ”top picks for you” interface. A superior set of log data would facilitate analysis across all recommendation interfaces in the system. Perhaps most importantly, we are attempting to find generalizable learning from a particular system (MovieLens + item-item CF) with a particular kind of item (movies). There is plenty of room for studying the differential rates of narrowing (or broadening) across media, and across algorithms. We hope our methods and results can be applied to inform the study of those domains.

Chapter 4

User Personality and User Satisfaction with Recommender Systems[1]

In the previous chapter, we found that a recommender system exposes users to narrower content over time. This narrowing effect may not fulfill user satisfaction with recommendations.

One approach to increase user satisfaction with recommendations is to generate recommendation lists with satisfactory levels of popularity, diversity, and serendipity (for example [28, 29, 42]). However, previous research assumes that all users have the same preferences for the levels of popularity, diversity, and serendipity of recommendation lists, or that recommendation algorithms can learn these preferences from users' ratings. We hypothesize that these assumptions are not true.

Furthermore, we investigate factors that can be integrated into recommendation algorithms to generate recommendations with the appropriate levels of diversity, popularity, and serendipity for individual users. User personality is a

[1] This work is under submission to ACM Transactions on Interactive Intelligent Systems (TiiS).

stable information source about individual user [16, 76], and has significant connections with users' tastes and interests [38, 77]. Therefore, we examine whether information about users' personalities can help recommendation systems deliver the appropriate levels of diversity, popularity and serendipity in recommendation lists, and increase users' satisfaction.

4.1 Research Questions

Therefore, we ask:

RQ1: How satisfied are users with the levels of diversity, popularity and serendipity of recommendation lists produced by rating-based recommendation algorithms?

RQ2: What is the relation between users' personality traits and their preferences for diversity, popularity, and serendipity?

RQ3: What is the relation between users' personality traits and their enjoyment of recommendations?

We conducted a study on MovieLens[2], a live movie recommendation system with more than 1,800 users. We showed each user a list of 12 personalized recommendations. We varied the levels of diversity, popularity, and serendipity of these lists. We then asked the users to rate how satisfied they were with the levels of diversity, popularity, and serendipity of these recommendation lists, and how much they would enjoy watching the movies in the lists. Finally, we assessed users' personality using the Ten-item Personality Inventory proposed by Gosling et al. [1].

Our contributions are three-fold. First, we demonstrate that user personality, in addition to user ratings, can provide good signals for recommender systems to deliver the satisfactory levels of recommendation diversity, popularity and serendipity to individual users. Our findings suggest that integrating user

[2] movielens.org

personality into recommendation algorithms could lead to increased user satisfaction. Second, we found that there are relations between user personality and user enjoyment with recommendations. Last but not least, we found that recommendation algorithms based on ratings alone could not always generate recommendation lists with desired levels of diversity, popularity, or serendipity.

4.2 Our Study

Our study examines the relations between users' personality types and their satisfaction with more or less diverse, popular and serendipitous movie recommendation lists. In this section, we describe our study in details.

4.2.1 User Experiment

We designed and launched a user experiment on MovieLens to collect data to answer our research questions. We ran the experiment from May 12th, 2015 to October 14th, 2015.

Eligible users

To participate in our study, a user must a) have an account with MovieLens, and b) rate at least 15 movies prior to his or her participation. We asked for the minimum of 15 movies because MovieLens traditionally required users to rate at least 15 movies before it could give the users any recommendation (see [41, 78]). We showed eligible users an invitation to participate in our study when users logged in to MovieLens. In total, 1,888 users participated in our study.

A between-subject experiment

Our user experiment is between-subject, with 10 conditions - 3 metrics (diversity, popularity, serendipity) × 3 different levels (high, medium, and low) and a control condition. Once a user accepted our invitation to participate in the study, we randomly assigned this user to one of the 10 conditions. Each user saw a list of 12 recommendations selected from the top 60 recommendations personalized for him or her. Thus, users saw different recommendation lists. We extracted the recommendation lists from the top 60 recommendations because our offline evaluation showed that the sixtieth recommendation was still a good recommendation with predicted ratings above 4 out of 5 starts for most users. We chose a between-subject design instead of a within-subject design because the our offline evaluation showed that with a within-subject design, each user would receive several overlapping recommendation lists. The high level of overlap among the recommendation lists that each user would receive affects our manipulation of diversity, popularity and serendipity. First, a repeated recommendation is no longer serendipitous to users after they have seen it several times. Second, users will perceive repeated recommendations as more popular. Third, users will perceive the recommendation lists as less diverse if they contain overlapping items.

How we varied the levels of diversity, popularity, and serendipity of recommendation lists

Choosing a random set of 12 recommendations from the top 60 takes a non-trivial amount of time because there are C_{12}^{60} possible ways to do so. Furthermore, it is time-consuming to choose twelve recommendations such that these recommendations are the most diverse or the least diverse. However, we want our users to see the recommendations within 30 seconds after they accepted the study invitation. Thus, we implemented a greedy search algorithm to select twelve recommendations out of the top sixty for diversity metric. The algorithms for popularity and serendipity metrics are straightforward. Below are the descriptions of these algorithms.

Diversity. We can estimate the diversity of a set by taking the average of all pairwise distances of all items in the set (see Ziegler et al.'s work [28]). We are interested in the tag-genome distances, which are measured using the tag-genome information space, similar to our work in chapter 3. More information about the tag-genome feature space can be found in the work of Vig et al. [79] or from GroupLens' dataset[3]. Since we want the diversity measure to be bounded by 0 and 1, we use cosine distance (i.e. 1 - cosine similarity) instead of Euclidean distance. Thus, given a set S of 12 movies, the diversity of S is defined as:

$$diversity\_score = \frac{1}{C_2^{12}} \times \sum_{\substack{m_i \in S \\ m_i < m_j}} \sum_{m_j \in S} cosine\_distance(m_i, m_j) \tag{4.1}$$

There are three different conditions for levels of diversity - high, medium and low. For the high-level condition, the algorithm selects twelve recommendations to form a list with a diversity score as high as possible (based on the above equation), and vice versa for the low-level condition. For the medium-level condition, the algorithm selects twelve recommendations to form a list with a diversity score in the middle.

We designed the algorithm for diversity metric with two parts. In the first part, the algorithm initializes a seeding list with a pair of movies. In the second part, the algorithm iteratively adds a movie into the seeding list until the number of movies in the list is 12. To create a seeding list, the algorithm selects a pair of movies out of C_2^{60} possible pairs. This pair must satisfy the following criteria. Among all the C_2^{60} possible pairs, if the condition is high-level, the pair has the highest diversity score; if the condition is low-level, the pair has the lowest diversity score; if the condition is medium-level, the pair has the median diversity score. The other fifty-eight movies form a remainder list. From the remainder list, the algorithm then subsequently selects a movie to add to the seeding list, so that, after adding, the new seeding list should have a) the highest possible diversity score if the condition is high-level, b) the lowest possible diversity score if the condition is low-level, c) the median diversity score if the condition is medium-level. After doing this, if the diversity score was not in the range of low, medium, or high, we exclude the user from our analyses (see section 4.5 for more details).

[3] http://grouplens.org/datasets/movielens/tag-genome/

Popularity. Popularity indicates how popular all the items in a set of recommendations are. We can estimate the popularity of an item by computing how many users have consumed (or rated) the item in the same system (based on [30]). In our study, we normalize the number of ratings per movie so that the approximated popularity ranges between 0% and 100%. Thus, given a set S of 12 movies where $num\_user\_rated(m_i)$ is the number of users rating movie m_i, the popularity score is defined as:

$$popularity\_score = \frac{1}{12} \times \sum_{m_i \in S} \frac{num\_user\_rated(m_i)}{number\_of\_distinct\_users} \times 100\% \qquad (4.2)$$

There are three different conditions for levels of popularity - high, medium and low. Thus, given the top 60 recommendations, our algorithm chooses 12 recommendations to form lists that have either high, or medium, or low level of popularity. First, for each movie m_i in the top 60 recommendations the algorithm computes the popularity score based on the above equation. The algorithm then sorts these movies from highest to lowest based on their popularity scores. The algorithm selects the top 12 movies for the high-level condition, the bottom 12 movies for the low-level condition, and the 25th to 36th movies for the medium-level condition.

Serendipity. Serendipity indicates how unexpected and interesting all the items in a set of recommendations are when compared with items users recently consumed (or rated). We can estimate the serendipity of a recommendation list by taking the average of all pairwise distances of between one item in the recommendation list, and one item the user has rated [29]. We are interested in tag-genome distances, similar to our calculation of diversity. Thus, we also use the tag-genome feature space to compute serendipity. Since we want the serendipity measure to be bounded by 0 and 1, we use cosine distance (i.e. 1 - cosine similarity). Furthermore, because users needed to rate at least 15 movies to participate in our study, we approximate serendipity in relation to the 15 movies a user rated most recently. Thus, for each user, given a set S of 12 movies and a set R of the 15 most-recently-rated movies, the serendipity score is defined as:

$$serendipity\_score = \frac{1}{12 * 15} \times \sum_{m_i \in S} \sum_{r_j \in R} cosine\_distance(m_i, r_j) \tag{4.3}$$

There are three different levels for serendipity metric - high, medium and low. Thus, given the top 60 recommendations, our algorithm chooses 12 recommendations to form lists that have either high, or medium, or low level of serendipity. First, for each movie m_i in the top 60 recommendations, the algorithm computes the serendipity score based on the above equation. The algorithm then sorts these movies from highest to lowest, based on their serendipity scores, and selects the top 12 movies for the high-level condition, the bottom 12 for the low-level condition, and the 25th to 36th movies for the medium-level condition.

Data analyses

While 1,888 users participated in our study, we only analyzed data from 1,635 users. We first excluded 253 users because of the following reasons:

- 203 users were assigned to the control condition.
- 47 users did not answer the questions assessing their personality.
- 3 users had fewer than 15 ratings prior to their participants in our study.

We then further excluded 416 users to whom the algorithms could not deliver the appropriate levels of diversity, popularity, or serendipity. Table 4.1 shows the number of users per condition in our data analyses. We discuss why we exclude these users in detail in section 4.5.

Table 4.1: The number of users who participated, and number of users who were included in analyses per condition.

Condition	**Block of recommendation property**	**Diversity**			**Popularity**			**Serendipity**		
	Level	Low	Medium	High	Low	Medium	High	Low	Medium	High
# of users who participated per condition		169	176	176	205	195	182	162	171	199
# of users included in data analyses per condition		150	147	171	151	135	158	114	102	92

We also note that:

- the correlations between the diversity and popularity levels, diversity and serendipity levels, and popularity and serendipity levels are -0.15, 0.17, and -0.31 respectively.
- the correlations between the user perceived diversity and popularity, user perceived diversity and serendipity, and user perceived popularity and serendipity are -0.05, 0.08, and -0.26 respectively.

4.2.2 Measures

Table 4.2: Questions to evaluate if users perceive the differences in recommendations with different levels of diversity, or of popularity or of serendipity

	Targeted block of recommendation properties	**Questions**	**How our users answer the question**
1	Diversity	How dissimilar are the movies in the list from each other?	On a 5-point scale: ● ● ● ● ● Very similar to each other Neutral Very dissimilar from Each other
2	Popularity	How popular are the movies in the list?	On a 5-point scale: ● ● ● ● ● Very obscure Neutral Very popular
3	Serendipity	How surprised are you to see these movies being recommended to you?	On a 5-point scale: ● ● ● ● ● Not surprised at all Neutral Very surprised .

After seeing the 12-item recommendation list, users answered two sets of questions.

Satisfaction with recommendation lists

In the first set of questions, we checked if users would notice the different levels of diversity, popularity, and serendipity of the movies in the recommendation list (see 4.2). Then we evaluate their satisfaction with the recommendations (see 4.3).

User Personality

The second set consists of 10 questions to assess their personality traits (see 4.4). We adopted these ten questions from the work of Gosling et al. [1] (p. 525). Each personality trait is measured by two questions of which one is the reversed

Table 4.3: Questions to evaluate users' satisfaction with the recommendation lists

	Targeted block of recommendation properties	Our interested factors about user satisfaction	Questions	How our users will answer the question?
1	Diversity	Diversity preference	Is the level of dissimilarity among the recommended movies right for you?	On a 5-point scale with the order as follow: Far too similar - A bit too similar - Just right for me - A bit too dissimilar - Far to dissimilar
2	Popularity	Popularity preference	Is the level of popularity of the recommended movies right for you?	On a 5-point scale with the order as follow: Far too obscure - A bit too obscure - Just right for me - A bit too popular - Far too popular.
3	Serendipity	Serendipity preference	Is the level of serendipity of the recommended movies right for you?	On a 5-point scale with the order as follow: Far too unsurprising - A bit too unsurprising - Just right - A bit too surprising - Far too surprising
4	All	Enjoyment	This list contains movies I think I would enjoy watching	On a 5-point likert scale from Strongly Disagree to Strongly Agree

Table 4.4: The ten questions to assess user personality adopted from Gosling et al. [1]. (*) denotes reversed questions.

#	Personality Trait	Question
		I see my self as:
1	Agreeableness	.... critical, quarrelsome. (*)
2		.... sympathetic, warm.
3	Conscientiousness	.... dependable, self-disciplined.
4		.... disorganized, careless. (*)
5	Emotional Stability	.... anxious, easily upset. (*)
6		.... calm, emotionally stable.
7	Extraversion	.... extraverted, enthusiastic.
8		.... reversed, quiet. (*)
9	Openness to experiences	.... open to new experiences, complex.
10		.... conventional, uncreative. (*)

of the other. Hence for each personality trait we first reverse the reversed question and take the average of the answers to the two questions. The average score per personality trait tells us the magnitude of this personality trait of a user.

Figure 4.1 shows the distributions of the five personality traits. Our users expressed that they were more introverted than extroverted. Hence, in this study, we use introversion instead of extraversion term where introversion is the reverse of extraversion. Likewise, since our users expressed that they were more emotionally stable, we use emotional stability instead of neuroticism term where emotional stability is the reverse of neuroticism.

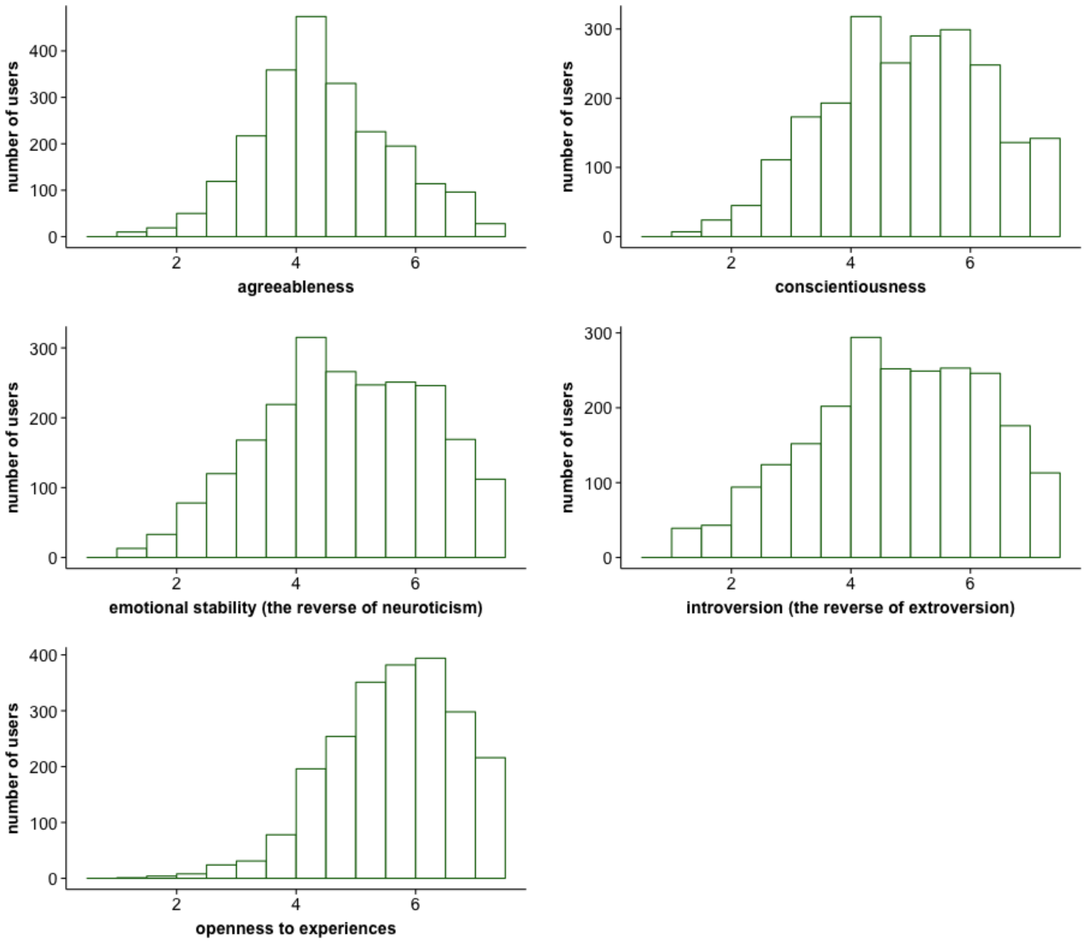


Figure 4.1: The distributions of the five personality traits

In this study, for each personality, we divide our users into two groups - a group with a high level (score greater than 4), and a group with a low level (score less than or equal to 4) of that personality trait.

Dividing our users into two groups (high vs. low) allow easy and understandable contrasts when we analyze the interaction between the personality traits and the 10 experimental conditions.

4.3 Results

4.3.1 The diversity, popularity, and serendipity of recommendation lists created by algorithms based only on user ratings did not satisfy most users

Table 4.5: Number of users who answered *Just Right* and number of users who would be satisfied with the levels of diversity, popularity and serendipity produced by rating-based only algorithms.

	Diversity			**Popularity**			**Serendipity**		
	High	Medium	Low	High	Medium	Low	High	Medium	Low
Number of users who answered Just Right	92	90	51	76	81	81	41	37	46
Number of the above users who would receive recommendations with their preferred levels of diversity, popularity, and serendipity created by rating-based only algorithms.	8	72	8	49	36	4	0	1	45

With the first research question, our goal is to estimate how satisfied users would be with the diversity, popularity, and serendipity levels of recommendations generated by default rating-based algorithms. Our data assesses users' satisfaction with the assigned category and level (e.g. high-diversity or low-serendipity). We, therefore, look at users who felt the assigned to level was *Just Right*.

These users are those who answered *Just Right* to the questions assessing their preferences for diversity, popularity and serendipity (i.e. the 1^{st}, 2^{nd}, and 3^{rd} questions in table 4.3). We report on how often these users would receive their preferred *Just Right*-level recommendations from rating-based algorithms. The reported frequencies is an estimate of user satisfaction with rating-based recommendation algorithms. We may overestimate the satisfaction because we have no evidence that users who were unsatisfied with the given level would have been happy at another level. We also may underestimate the satisfaction because users who are happy at their assigned level might be happy at multiple levels.

Diversity. For diversity, 233 out of 468 users reported *Just Right* satisfaction (92/171, 90/147, 51/150). Of these, only 88 would have received the preferred level of diversity from rating-based algorithms (8/92, 72/90, 8/51). This ratio suggests that rating-based algorithms could satisfy users with medium-diversity preferences 80% of the time. However, for many users who were happy with high- or low-diversity, rating-based algorithms could satisfy only 9%-16% of the time.

Popularity. We carry out the same calculations as above and found that 89 out of 238 users would have received the preferred level of popularity from rating-based algorithms (49/76, 36/81, 4/81). This fraction suggests that rating-based algorithms could satisfy users with high-popularity preferences 65% of the time, and users with medium-popularity preference 44% of the time. Although about one-third of 238 users preferred low-popularity, rating-based algorithms could satisfy their satisfaction only 5% of the time.

Serendipity. With the same calculations, we found that 46 out of 124 users would have received the preferred level of serendipity from rating-based algorithms (0/41, 1/37, 45/46). This ratio suggests that rating-based algorithms guarantee to satisfy users have low-serendipity preferences (98% of the time) and that the algorithms are certain to fail users who have high- and medium-serendipity preferences (0% of the time).

Table 4.5 displays the number of users who are satisfied with the levels of diversity, popularity, and serendipity of recommendation lists created by rating-based algorithms.

4.3.2 User personality and user preference for diversity, popularity and serendipity.

The second research question ask if there is a correlation between users' personality traits and their preferences for diversity, popularity, and serendipity? To answer this question, we tested the interaction effect between each personality trait and the levels of recommendation property diversity, popularity, and serendipity. Significant interaction effects suggest that users with different levels of personality trait have different preferences for how diverse, popular, and serendipitous a recommendation list should be. We used ordinal logistic regression[4] with the following model.

$$folded.response \sim personality.level * score^5 \tag{4.4}$$

Ordinal logistic regression is appropriate in this case because the dependent variable is an ordered categorical variable, and we are interested in interpreting the probabilities of users choosing ordinal answers.

For all targeted block of recommendation properties, the dependent variable is users' answers for 1^{st}, 2^{nd},3^{rd} questions assessing their preferences for diversity, popularity, and serendipity respectively (as shown in table 4.3). Because these questions have the most satisfactory option in the middle of the scales, we transformed the scales into new variables as follows. For diversity, we recoded *Far too similar* and *Far to dissimilar* options to *far* , *A bit too similar* and *A bit too dissimilar* options to *close*, and *Just Right* to *Just Right*. For popularity, we recoded *Far too obscure* and *Far too popular* to *far*, *A bit too obscure* and *A bit too popular* to *close*, and *Just Right* to *Just Right*. For serendipity, we recoded *Far too unsurprising* and *Far too surprising* to *far*, *A bit too unsurprising* and *A bit too surprising* to *close*, and *Just Right* to *Just Right*.

The independent variable scores are the diversity, popularity, and serendipity scores computed based on equations 4.1, 4.3, and 4.2 respectively for diversity, popularity and serendipity.

Below, we describe the results for each recommendation property (diversity, popularity, and serendipity).

Diversity. We found only the interaction effect between high-introversion and

[4] using polr in R.

[5] Although in our user experiment we have three categorical levels for diversity, popularity, and serendipity, we decide to analyze the data based on continuous variable represented the diversity, serendipity and popularity of recommendation lists. It is because the distributions for diversity, popularity and serendipity are either skewed or highly overlapped as shown in 4.5. Thus using continuous variables makes our analyses independent from the distributions and easily duplicated.

diversity score significant (p.value of interaction effect = 0.05[6],[7]). This significant effect suggests that more high-introversion users preferred diverse recommendations than similar ones. Specifically, the effect suggests that 0.1 (10%) increases in diversity correspond to 1.12 increases in the odds of high-introversion users choosing *Just Right*. However, increases in diversity do not significantly change the odds of low-introversion users choosing *Just Right*. We visualize the changes in odds of choosing *Just Right* for high- and low-introversion users in figure 4.2.

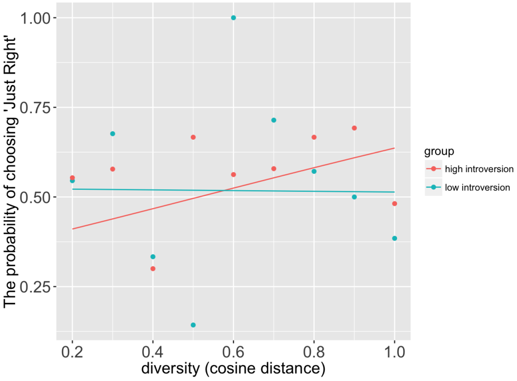


Figure 4.2: The probabilities of high and low introverted users choosing *Just Right* with different diversity quantity.

Popularity. We found that only the interaction effect between introversion and popularity significant (the interaction effect p.value = 0.02)[8]. We also found only the main effect of conscientiousness significant (p.value = 0.02).

The interaction effect suggests that more low-introversion users preferred popular recommendations than non-popular ones, but more high-introversion users preferred non-popular recommendations than popular recommendations A 10% increases in popularity score corresponds to a 1.01 increase in the odds of low

[6] The statistical significance test reported is of the interaction effect only; i.e., of the difference between the odds-ratios of high- and low–introversion users.

[7] The model is reported in section 4.6 table table 4.6.

[8] The model is reported in section 4.6 table 4.7

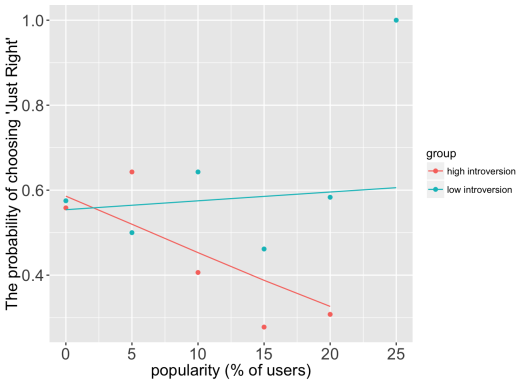


Figure 4.3: The probabilities of high and low introverted users choosing *Just Right* with different popularity quantity.

introversion users choosing *Just Right* , but a 0.55 decrease in the odds of high-introversion users choosing *Just Right*. We visualize the changes in the odds of high- and low- introversion users choosing *Just Right* in figure 4.3

The main effect of conscientiousness suggests that no matter how popular the recommendation lists, low-conscientiousness users were more likely to choose *Just Right* (60% of the time) than high-conscientiousness users (48% of the time)[9].

Serendipity. We found three significant interaction effects. These effects are between serendipity and conscientiousness trait (the interaction effect p.value = 0.05), between serendipity and introversion trait (the interaction effect p.value = 0.00), and between serendipity and openness trait (the interaction effect p.value =0.03).

We found that more high-conscientiousness users preferred low than high serendipitous recommendations, but more low-conscientiousness users preferred high than low serendipitous recommendations. More high-introversion users preferred high than low serendipitous recommendations, but more low-introversion users preferred low than high serendipitous recommendations. More low-openness

[9] The model is reported in section 4.6 table 4.11

users preferred low serendipitous than high serendipitous recommendations. High-openness users did not have any clear preference for the level of serendipity.[10] These significant interaction effects suggest that 0.1 (10%) increases in serendipity corresponds to:

- * 1.06 increases in the odds of low-conscientiousness users, but 0.86 decreases in the odds of high-conscientiousness users choosing *Just Right* .
- * 0.82 decreases in the odds of low-introversion users, but 1.03 increases in the odds of high-introversion users choosing *Just Right*.
- * 0.76 decreases in the odds of low-openness users choosing *Just Right*. However, the change in the odds of high-openness users choosing *Just Right* is negligible .

Figure 4.4 displays the changes in the odds of users choosing *Just Right* for different personality traits and levels of serendipity

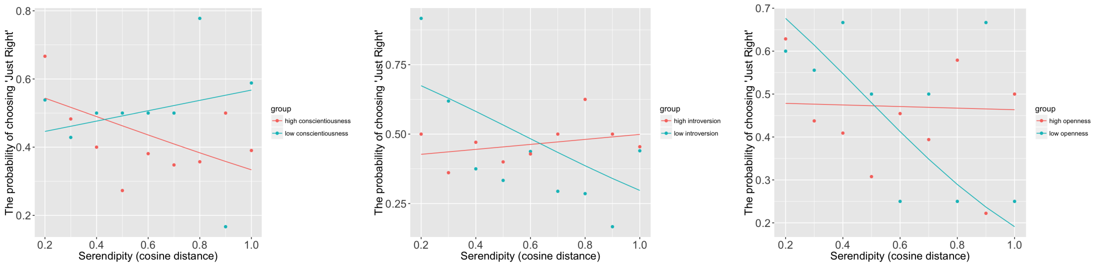


Figure 4.4: The probabilities of users choosing *Just Right* with different serendipity quantity. Left: conscientiousness; Middle: introversion; Right: openness

User satisfaction and user enjoyment

In summary, we found that introversion, conscientiousness, and openness personality traits contain strong signals about user preference for diversity, popularity and serendipity. However, delivering individually preferred amounts of diversity,

[10] These models are reported in section 4.6 tables 4.8, 4.9,4.10.

popularity or serendipity does not necessarily leads to the increase in user enjoyment. In fact, we found correlations, though not big, between user satisfaction with recommendation diversity, popularity and serendipity The correlation between user satisfaction with the diversity of the recommendation list and the user enjoyment of the movies in the list is 0.21. The correlation is 0.24 between users' satisfaction with the popularity of the list and their enjoyment, and 0.12 between users' satisfaction of the serendipity of the list and their enjoyment. Therefore, in the next section, we investigate the relationship between user personality and user enjoyment for recommendation lists with different levels of diversity, popularity and serendipity.

4.3.3 User personality and user enjoyment for recommendations with different levels of diversity, popularity and serendipity

The third research question asks if there is a relation between users' personality traits and their enjoyments of the recommendations?
For each of the experimental conditions, we test if the interaction effects of personality traits and the self-reported enjoyment of the movies in the recommendation list were significant. We employed ordinal logistic regression with the following model:

$$response \sim personality.level * score \tag{4.5}$$

where the dependent variable is the self-reported enjoyment of the recommendations (users' answers to the 4^{th} question shown in table 4.3. For ease of interpretation, we recoded users' answers to this question from 5-point Likert scale ranging from *Strongly Disagree* to *Strongly Agree* into three categories: *Disagree, Neutral, and Agree.*

Diversity. We found only the interaction effect between diversity and introversion trait significant (the p-value of the interaction effect is 0.02)[11]. The significant effect means low-introversion users did not enjoy the recommendations in a diverse list as much as they enjoyed the recommendations in a less diverse list. Specifically, 0.1 (10%) increases in the diversity of the list correspond to 0.86 decreases in the odds of low-introversion users expressing that they would enjoy the recommendations in the list. Meanwhile, high-introversion would equally enjoy recommendations from a diverse list or an unvaried list. We visualize the changes in the odds of enjoying the recommendations for high- and low-introversion users in figure 4.5 below.

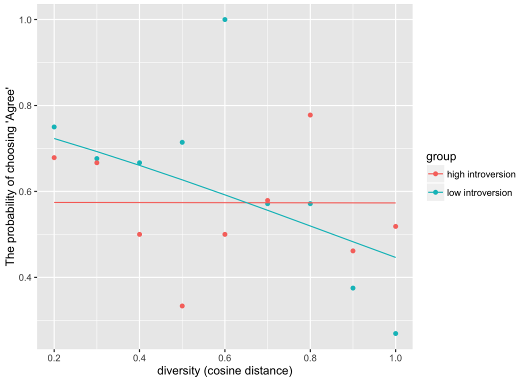


Figure 4.5: The probabilities of high and low introverted users reported that they would enjoy the recommendations with different diversity quantity.

Popularity. We found only the interaction effect between high-emotional-stability and enjoyment significant (the interaction effect p-value is 0.02)[12]. The interaction effect shows that high-emotional-stability users were more likely to enjoy recommendations from a popular list than low-emotional-stability users. Specifically, it suggests that 10% of increases in the popularity of the list correspond to two times (2x) increases in the odds of high-emotional-stability users self-reporting that they would enjoy the recommendations in the list. However, low-emotional-stability

[11] The model is reported in section 4.6 table 4.12.

[12] The model is reported in section 4.6 table 4.13.

users would equally enjoy recommendations from a popular or a less popular list. Figure 4.6 shows the changes in odds of high- and low-emotional-stability users.

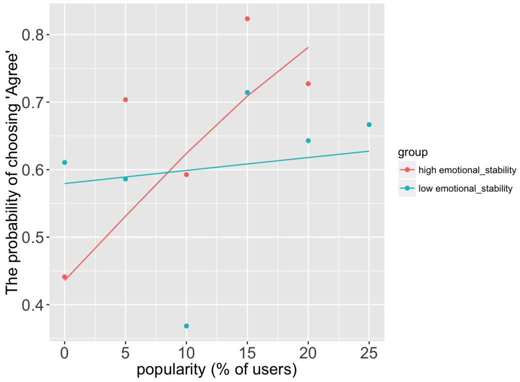


Figure 4.6: The probabilities of high and emotional stability users reported that they would enjoy the recommendations with different popularity quantity.

Serendipity. We found no significant interaction effect between personality traits and serendipity scores (all p.values > 0.1).

4.4 Conclusion

In summary, we found that users have different preferences for diversity, popularity, and serendipity. We also found that rating-based algorithms do not consistently deliver the diversity, popularity, and serendipity levels preferred by individual users. On the other hand, users' personality traits provide important signals to predict these individual preferences and overall recommendation enjoyment.

Our findings demonstrate how recommender systems can integrate user personality information into their recommendation-generation frameworks. Prior work concentrated on incorporating the information to improve prediction accuracy. For example Hu et al. [15] blended the personality information into collaborative filtering frameworks and saw improvements in MAE and ROC sensitivity metrics. Tkalcic et al. [80] showed that using personality-based user-similarity in a collaborative filtering algorithm does not diminish the prediction accuracy when compared with rating-based algorithms. Other researchers focused on different aspects of user experience such as acceptance issues of personality-based recommender systems [81]. Our work contributes to the understanding of improving the overall user experience from the angle of user preference for diversity, popularity, and serendipity.

While it is not common today for a recommender system to know personality traits, and we are not advocating giving a personality test before signing up. There is an increasing amount of user data available through federated sites, including sites that claim to have personality information. It is reasonable to believe that future commercial sites (e.g., Amazon, Facebook or Pinterest) will have access to personality assessment data for at least some of their users. Moreover, it's not necessary to always assess user personality of all users via questionnaires. Prior work showed that we can reliably predict users' personality based on their online footprint. Quercia et al. [82] and Bachrach et al. [83] demonstrated that we can predict well users' personality based on their Twitter or Facebook profiles. Youyou et al. [84] went further to prove that well-trained algorithms can do a better job than humans at predicting user personality.

Our work also confirms that satisfaction with recommendations is a property of the entire set of recommendations and not only of the individual items recommended (e.g. [4, 6]). This satisfaction, in turns, has strong connections with the other properties of the recommendation sets such as diversity and serendipity. Although popularity is a property of individual recommendation, our study proves that when considered as a property of a set, popularity also has a strong connection with user satisfaction. We can infer the directions of these connections via users' personality traits.

We think it is important to continue exploring the ways in which personality data can be used to improve user experience with recommender systems. One direction for future work is to take the n-dimensional matrix factorization approach proposed by Karatzoglou et al. [85] and treat the effects of personality traits that we found in our study as new dimensions. An alternative is to consider a Bayesian personalized ranking (e.g. Rendle et al.'s work [86]) and treat the effects of personality as prior belief. We also believe that future work should investigate the effects of all five personality traits together, as each trait plays an equal role in influencing users' decisions.

Our study is not without limitation. The choice of using a between-subject design for our experiment, even after careful evaluations, poses several challenges for our analyses. First, we have limited data per user, and our data does not have contrast in users' preferences for different levels of diversity, popularity and serendipity. Second, we only examine one recommendation property at a time. We also cannot calibrate our users' desired levels of diversity, popularity, and serendipity. Moreover, we had to exclude users for whom randomly-assigned levels of diversity, popularity, and serendipity could not be achieved. Furthermore, the statistical models that we used do not perfectly fit our observed data. This may be because we used only ten questions to approximate user personality and that the distributions of user personality questions are skewed.

We also only examine user satisfaction and user personality in the movie domain with a specific recommender system - MovieLens. In the future, we would like to conduct similar research in different domains - such as books or news. Duplicating this study in different domains will enhance our knowledge of user satisfaction with recommendations.

4.5 Reason to exclude some users from the analyses.

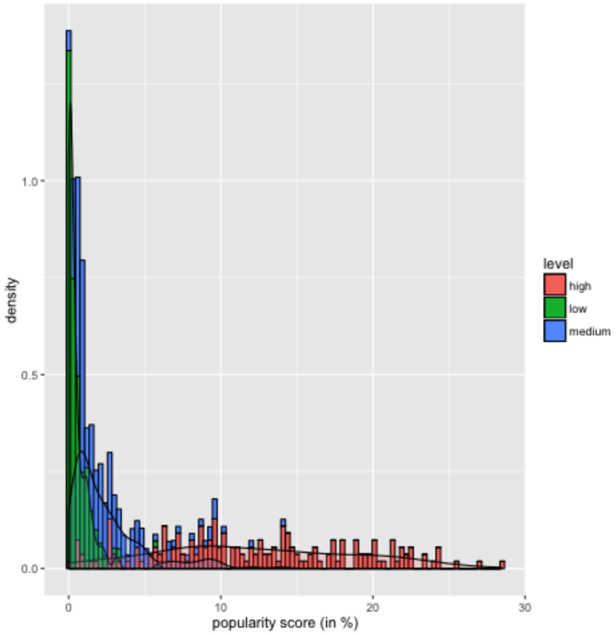


Figure 4.7: The distributions of popularity scores. The score ranges from 0 to 1. Red indicates high level, blue medium level and green low level. There are three curves indicating the three peaks of high, medium and low levels of popularity

Before explaining why we exclude some users from our analyses, we discuss the results of our manipulation-check analyses. The analyses contain insights for the exclusion.

Our manipulation-check analyses show that

- for diversity metric, users perceived the differences in recommendations of high and low levels (p.value = 0.000), and of medium and low levels (p.value = 0.000). Users did not perceived the differences in recommendations of medium and high levels (p.value = 0.255).

- for popularity metric, users perceived the differences in recommendations

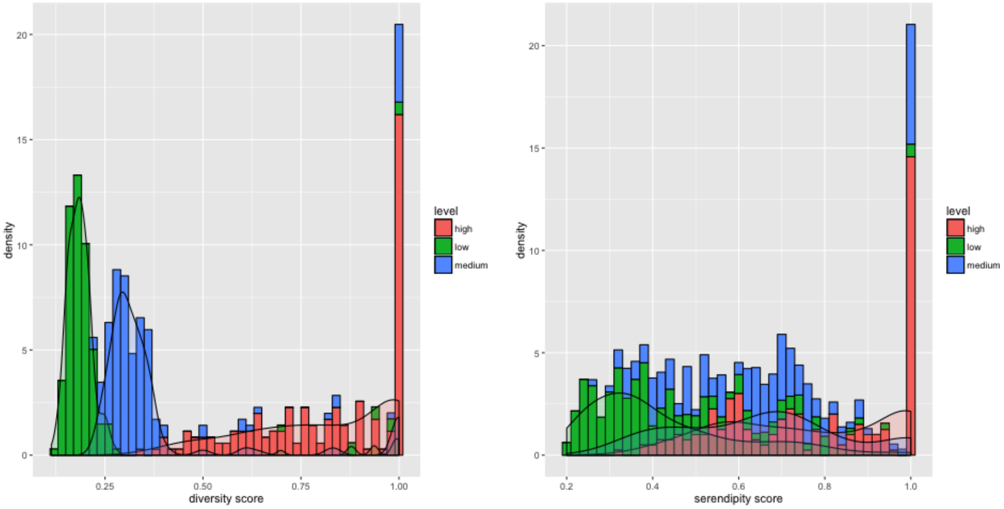


Figure 4.8: The distributions of diversity scores (on the left) and serendipity scores (on the right). Both scores range from 0 to 1. Red indicates high levels, blue medium levels and green low levels. In each figure, there are three curves indicating the three peaks of high, medium, and low levels of divesity (left figure) or of serendipity (right figure)

of high and low levels ($\text{p.value} = 0.000$), in recommendations of medium and high levels ($\text{p.value} = 0.000$). Users also perceived the differences in recommendations of medium and low levels (with marginally $\text{p.value} = 0.057$).

- for serendipity metric, users perceived the differences in recommendations of high and low levels ($\text{p.value} = 0.035$), in recommendations of medium and low levels ($\text{p.value} = 0.021$). Users did not perceive the differences in recommendations with medium and high levels ($\text{p.value} = 0.956$)

We plot out the distributions of diversity, popularity, and serendipity to investigate why in some level comparisons (e.g. high vs. medium diversity level) users did not perceive the differences. Figures 4.7 and 4.8 show these distributions. In each distribution, we observe overlapping regions of high, medium and low levels. These overlappings make users not perceive differences in some recommendations with different levels.

In this study, we want to examine the recommendation experience of users who perceived the differences in the diversity, popularity, or serendipity quantities per level. Thus, we remove from our analyses users to whom the recommendations cannot deliver to the appropriate quantities of diversity, popularity, and serendipity. Our process to remove users from the analyses is as follow.

Diversity. From figure 4.8 (left), there are clear cut-off points for low, medium and high levels of diversity at 0.24 and 0.41. Thus, we analyze 150/169 users who were assigned to the low-level with diversity score less than 0.24, 147/176 users assigned to the medium-level with diversity score from 0.24 to 0.41, and 171/176 users assigned to high level with the score greater than 0.41.

Popularity. From figure 4.7, there are clear cut-off points for low, medium and high levels at 0.6% and 5.0%. Thus, we analyze 151/205 users assigned to low-level with popularity score from 0% to less than 0.6%, 135/195 users assigned to medium-level with popularity score from 0.6% to 5.0%, and 158/182 users assigned with high-level with popularity score greater than 5.0%.

Serendipity. From figure 4.8 (right), there are clear cut-off points for low, medium and high levels of serendipity at 0.46 and 0.80. Thus, we analyze 114/162 users who were assigned to low-level with serendipity score less than 0.46, 102/171 users assigned to medium-level with serendipity score from 0.46 to 0.80, 92/199 users with serendipity score greater than 0.80.

4.6 Tables for fitted models

Table 4.6: Ordinal regression results for diversity preferences of introversion personality trait.

Variables	Logistic Coefficient	Standard error	95% CI	p value	Increase in Odds Ratio
Personality level					
High	-0.69	0.33	[-1.34, -0.03]	0.04	0.50
Low	Ref				
Diversity quantity					
Score (cosine distance)	-0.04	0.46	[-0.95, 0.87]	0.90	0.96
Interaction effect					
high*score	1.19	0.62	[-0.02, 2.41]	0.05	3.29
low*score	Ref				

Goodness of fit: Deviance chi-square = 790, df = 463 (p = 0.00).

Table 4.7: Ordinal regression results for popularity preferences of introversion personality trait.

Variables	Logistic Coefficient	Standard error	95% CI	p value	Increase in Odds Ratio
Personality level					
High	0.13	0.25	[-0.35, 0.61]	0.50	1.14
Low	Ref				
Popularity quantity					
Score (% of users rated)	0.001	0.02	[-0.03, 0.05]	0.60	1.0
Interaction effect					
high*score	-0.06	0.03	[-0.12, -0.01]	0.02	0.93
low*score	Ref				

Goodness of fit: Deviance chi-square = 742, df = 439 (p = 0.00).

Table 4.8: Ordinal regression results for serendipity preferences of conscientiousness personality trait.

Variables	**Logistic Coefficient**	**Standard error**	**95% CI**	**p value**	**Increase in Odds Ratio**
Personality level					
High	0.73	0.58	[-0.41, 1.87]	0.20	2.01
Low	Ref				
Serendipity quantity					
Score (cosine distance)	0.60	0.69	[-0.75, 1.97]	0.30	1.84
Interaction effect					
high*score	-1.70	0.87	[-3.41, 0.00]	0.05	0.18
low*score	Ref				

Goodness of fit: Deviance chi-square = 536, df = 303 (p = 0.00).

Table 4.9: Ordinal regression results for serendipity preferences of introversion personality trait.

Variables	**Logistic Coefficient**	**Standard error**	**95% CI**	**p value**	**Increase in Odds Ratio**
Personality level					
High	-1.49	0.61	[-2.69, -0.29]	0.02	0.23
Low	Ref				
Serendipity quantity					
Score (cosine distance)	-1.99	0.71	[-3.42, -0.61]	0.00	0.14
Interaction effect					
high*score	2.34	0.88	[0.62, 4.11]	0.00	10.44
low*score	Ref				

Goodness of fit: Deviance chi-square = 534, df = 303 (p = 0.00).

Table 4.10: Ordinal regression results for serendipity preferences of openness personality trait.

Variables	**Logistic Coefficient**	**Standard error**	**95% CI**	**p value**	**Increase in Odds Ratio**
Personality level					
High	-1.35	0.80	[-2.92, 0.21]	0.08	0.25
Low	Ref				
Serendipity quantity					
Score (cosine distance)	-2.73	1.08	[-4.90, -0.62]	0.01	0.07
Interaction effect					
high*score	2.66	1.17	[0.35, 4.95]	0.02	14.17
low*score	Ref				

Goodness of fit: Deviance chi-square = 535, df = 303 (p = 0.00).

Table 4.11: Ordinal regression results for popularity preferences of conscientiousness personality trait (main effects).

Variables	Logistic Coefficient	Standard error	95% CI	p value	Increase in Odds Ratio
Personality level					
High	-0.46	0.19	[-0.84, 0.08]	0.02	0.63
Low	Ref				

Table 4.12: Ordinal regression results for users' self-reported enjoyment with different diversity quantities for introversion personality trait.

Variables	Logistic Coefficient	Standard error	95% CI	p value	Increase in Odds Ratio
Personality level					
High	-0.95	0.35	[-1.65, -0.27]	0.00	0.39
Low	Ref				
Diversity quantity					
Score (cosine distance)	-1.47	0.46	[-2.37, -0.58]	0.00	0.23
Interaction effect					
high*score	1.46	0.60	[0.28, 2.64]	0.02	4.32
low*score	Ref				

Goodness of fit: Deviance chi-square = 877, df = 463 (p = 0.00).

Table 4.13: Ordinal regression results for users' self-reported enjoyment with different popularity quantities for emotional stability personality trait.

Variables	Logistic Coefficient	Standard error	95% CI	p value	Increase in Odds Ratio
Personality level					
High	-0.57	0.24	[-1.06, -0.11]	0.02	0.56
Low	Ref				
Diversity quantity					
Score (cosine distance)	0.00	0.02	[-0.03, 0.05]	0.60	1.01
Interaction effect					
high*score	0.07	0.03	[0.01, 0.13]	0.02	1.01
low*score	Ref				

Goodness of fit: Deviance chi-square = 866, df = 439 (p = 0.00).

Chapter 5

The relations between user satisfaction measured with survey questionnaires and users' consumption measured with logged data.

Recommender system researchers are progressively attempting to evaluate not only the accuracy of recommendation algorithms, but also the user satisfaction with recommendations produced by the algorithms. In the research community, most such evaluations are based on user survey where users specifically answer questions evaluating their enjoyment, satisfaction, trust, and many other factors. In the industry by contrast, researchers and practitioners usually evaluate user satisfaction via post-hoc analyses about users' consumptions with user-behavior logged data.

Although both approaches can be used independently from each other to assess different aspects of the user satisfaction with recommendations, understanding how these findings from both approaches are related is important. Particularly, understanding whether and how the measured user satisfaction via survey questions relates to the measured user consumption via logged data could help researchers and practitioners prioritize what to measure about user satisfaction with survey questionnaires. In this work, we investigate the relations between the user satisfaction measured with survey questionnaires and the users' consumption measured with logged data.

Thus, we carry out a user study in which we employed both survey questionnaire and logged data analysis. With survey questionnaire we measure user satisfaction with recommendations along dimensions of perceived personalization, perceived enjoyment, and how easily they think of other movies they would enjoy more. We also evaluate their satisfaction with the three recommendation properties: diversity, popularity, and serendipity. Moreover, we capture the three properties of recommendations and recommendation lists at the time users answered the survey questions. Those are diversity, popularity, and serendipity of recommendations and recommendation lists. We monitor the users' consumptions over a period of three months, and analyze the relations between the users' answers for our survey questions and their consumption.

5.1 Research Questions

With the collected data, we set out to answer two research questions:

RQ1: Whether and how the user self-reported satisfaction relates to the inferred user recommendation consumption?

RQ2: Whether and how the three properties of recommendations and recommendation lists relate to the inferred user recommendation consumption?

Our findings, in brief, are the followings. We find that the user self-reported satisfaction with recommendation popularity and serendipity is highly related to the subsequent user recommendation consumption. Interestingly, we do not find any evidence suggesting that the perceived enjoyment, and perceived personalization correspond to the subsequent user recommendation consumption. Thus, because not all metrics evaluating the user experience correspond to the user subsequent recommendation consumption, we conclude that combining survey-questionnaires with logged data analyses may provide a more complete picture about the user experience with recommender systems.

5.2 Our study & our measures

In this section, we describe our study in detail. We also discuss what we collect and measure at the time of users answering our survey questions, and how we infer user recommendation consumption (i.e. the inferred user satisfaction).

5.2.1 Our Study

The purpose of our study is to investigate the relations between user satisfaction measured with survey questionnaires and their subsequent consumption. To that end, we deploy a user study on MovieLens, a well-known research recommender system. We then monitor the consumptions of those who participate in our study for a three-month period.

Eligibility. To participate in our study, a user had to have a MovieLens account and had rated at least fifteen movies. We asked for a minimum of fifteen movies because traditionally users needed to rate this number of movies to pass the cold-start process [41, 78].

A survey questionnaire. The first part of our study consists of a survey to measure the user self-reported satisfaction. Once users accepted our invitation to participate in our study, they were presented with a list of the top twelve personalized recommendations, and a set of questions to evaluate their satisfaction with the levels of diversity, popularity, or serendipity of the recommendation in

Table 5.1: The questions in the survey to measure the subjective factors about user satisfaction.

#	Subjective factors		Question	How users answered the question.
1	**Perceived**	enjoyment	This list contains movies I think I would enjoy watching	On a 5-point likert scale from Strongly Disagree to Strongly Agree
2		personalization	This list is personalized for me	On a 5-point likert scale from Strongly Disagree to Strongly Agree
3		bestness	I can easily think of movies I rather watch than those on this list.	On a 5-point likert scale from Strongly Disagree to Strongly Agree
4	**User satisfaction with**	recommendation diversity	Is the level of dissimilarity among the recommended movies right for you?	On a 5-point scale with the order as follows: Far to similar - A bit too similar - Just right for me - A bit too dissimilar - Far too dissimilar.
5		recommendation popularity	Is the level of popularity of the recommended movies right for you?	On a 5-point scale with the order as follows: Far too obscure - A bit too obscure - Just right for me - A bit too popular - Far too popular.
6		recommendation serendipity	Is the level of serendipity about the movies right for you?	On a 5-point scale with the order as follows: Far too unsurprising - A bit too unsurprising - Just right for me - A bit too surprising - Far too surprising

the list. These questions are listed in table 5.1. The top twelve personalized recommendations were generated by MovieLens' recommendation engines without any manipulation.

A longitudinal data analysis. In the second part of our study, we log user recommendation consumption over the course of three months. We infer user recommendation consumption by looking at the user rating behaviors (i.e. if a user rated a recommendation, we assume he or she consumed the recommendation). Knowing if users actually take recommendations is a hard problem since we cannot ask users to confirm. Therefore, we apply the method we used in chapter 3. That is we start logging user rating behaviors three hours and stop logging three months after users answer our survey questions. We visualize our monitoring period in figure 5.1.

5.2.2 What we measure

We use six metrics for user self-reported satisfaction, three recommendation- and recommendation-list- properties that can influence the user satisfaction, and the inferred user recommendation consumption. We explain these measures as follows.

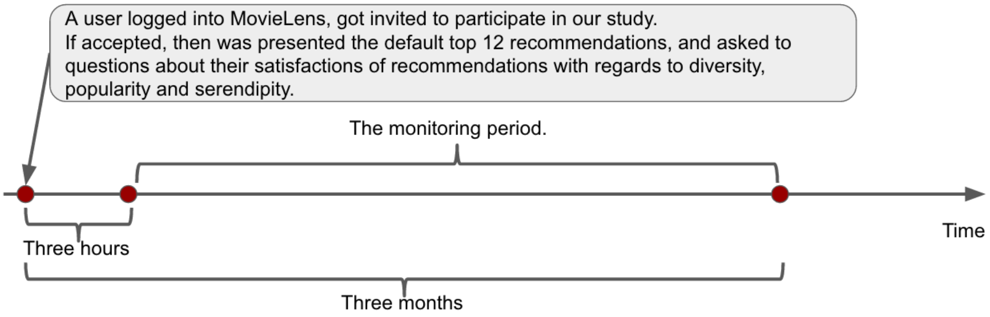

A user logged into MovieLens, got invited to participate in our study.
If accepted, then was presented the default top 12 recommendations, and asked to questions about their satisfactions of recommendations with regards to diversity, popularity and serendipity.
The monitoring period.
Time
Three hours
Three months

Figure 5.1: Our monitoring period.

User self-reported satisfaction.

With the questions in table 5.1, we measure the following self-reported satisfaction: *perceived enjoyment* (the 1^{st}*question*), *perceived personalization* (the 2^{nd} question), *how easily they think of other movies they rather watch* (3^{rd}*question*), and user satisfaction with recommendation diversity, popularity, and serendipity (the 4^{th}, 5^{th}, and 6^{th} question respectively).

Properties of recommendations and recommendation lists

In this research, we focus on three properties of recommendations and recommendation lists that prior research shows to influence user satisfaction with recommendations. Those are recommendation popularity, diversity, and serendipity [28, 29, 42].

Because we use a tag-genome information space to measure diversity and serendipity of the recommendations, we will describe the information space first before discussing how we measure all the objective factors.

Tag-genome information space. The tag-genome information space is a set of descriptive tags that describe the content of movies. In this information space, each pair of a tag t and a movie m is associated with a relevant score rel. These relevant scores are arithmetically computed based on the inputs of MovieLens users who determined how good the tag describes the movie. A relevant score is ranged from 0 to 1 where 0 means that a tag does not describe a movie at all, and 1 means that a tag describes a movie well. This tag-genome information space is built by Vig et al. [79] to help users navigate through thousands of movies while looking for a movie that is very similar in all, but one, content-dimensions with an anchoring movie.

The tag-genome information space we use in this study consists of 1,100 tags describing 10,000 movies, and can be downloaded from the GroupLens dataset[1]. Please refer to section 1.3.2 for more details.

Diversity. We measure recommendation diversity by approximating the overall diversity of a recommendation list with tag-genome. For each recommendation list, we compute a diversity score to approximate how diverse the recommendation list is. A diversity score is computed by taking the average of all pairwise cosine distances of the 12 movies in the list. This approach of approximating list diversity is widely used in recommender system literature [28, 41]. Thus, given a set S of twelve movies, the diversity of S is defined as:

$$diversity\_score = \frac{1}{C_2^{12}} \times \sum_{\substack{m_i \in S \\ i<j}} \sum_{m_j \in S} cosine\_distance(m_i, m_j) \tag{5.1}$$

where C_2^{12} is $\frac{12!}{(10!*2!)}$.

As mentioned, these cosine distances are not based on the rating vectors, but on the tag-genome information space described above. Hence, a computed diversity score indicates how diverse in content a recommendation list is.

Popularity. We measure recommendation popularity by approximating the overall popularity of a recommendation list. The overall popularity of a list is the average recommendation popularity of all recommendation in the list. A recommendation popularity can be approximated by computing how many users have consumed (or rated) the item in the same system [30]. In our study, we normalize the number of ratings per movie so that the approximated popularity is bounded by 0 % to 100% where 0% means overall the movies in a list are rarely

[1] http://grouplens.org/datasets/movielens/tag-genome/

rated by MovieLens users and 100% means the movies in a list are all rated by MovieLens users. Thus, given a set S of 12 movies with $num\_users\_rated(m_i)$ is the number of distinct users rating movie m_i, the popularity score is defined as:

$$popularity\_score = \frac{1}{12} \times \sum_{m_i \in S} \frac{num\_users\_rated_(m_i)}{number\_of\_distinct\_users} \times 100\% \tag{5.2}$$

Serendipity. We measure recommendation serendipity by approximating how unexpected but interesting a recommendation list is compared with the user's most recently rated fifteen movies. Serendipity is computed by taking the average of all pairwise distances of two items - one is from the recommendation list and the other is from the most fifteen recently rated movies. This approach was used to approximate how serendipitous a music recommendation list is [29].

Thus, given a set S of twelve movies and a set R of the most recently rated fifteen movies by a user, the serendipity score is defined as:

$$serendipity\_score = \frac{1}{12 * 15} \times \sum_{m_i \in S} \sum_{r_j \in R} cosine\_distance(m_i, r_j) \tag{5.3}$$

Like diversity, these cosine distances are computed based on the tag-genome information space described above, hence indicate how serendipitous from the usual users' consumption the content recommendation list is from the most recently rated fifteen movies.

Inferred user consumption with logged data.

From a longitudinal data analysis, we first remove all users who did not rate any movie during the monitoring period. This is to avoid noise added by uncommitted users who did not really use MovieLens recommender system. Those users might be in the phase of exploring MovieLens while participating in our study and they never came back to use MovieLens.

For users who rated at least one movie during the monitoring period, if one of their rated movies was in the top twelve recommendations presented to them, we classify these users as users as a group that took at least one recommendation. If none of their rated movies was in the top twelve recommendations presented to them, we classify these users as a group that did not take any recommendation.

5.2.3 The collected data

We run our user study on Movielens from May 12^{th} 2015 to October 14^{th} 2015, collecting data of 203 users who were presented with the top twelve recommendations and answered the questions in table 5.1. After answering the questions, the users could always see the top twelve recommendations during the monitoring period. Of these 203 users, we only analyze the data of 116 users who rated at least one movie during the monitoring period. This is as mentioned to avoid noise added by uncommitted users who did not really use MovieLens recommender system.

5.3 Results

We present our results for whether and how user self-reported satisfaction and the properties of recommendations and recommendation lists relate to the inferred user recommendation consumption.

5.3.1 User self-reported satisfaction

In this section, we seek to answer the first research question:

RQ1: Whether and how the user self-reported satisfaction relates to the inferred user recommendation consumption?

Self-reported enjoyment. The first factor we look at is self-reported enjoyment. We seek to answer whether the self-reported enjoyment relates to the subsequent user recommendation consumption. To that end, we looked at how 116 users answered the 1^{st} question in table 5.1. Of 116 users, 78 took at least one recommendation and 38 who did not take any recommendation. We then look at how their answers distribute across the answering scale respectively for groups of

78 and 38 users. Table 5.2 shows the two distributions. We plot the percentage of users who took at least a recommendation at per answer in figure 5.2.

Table 5.2: The number of users per answer for enjoyment and whether they took at least a recommendation from the list.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Take at least one recommendation	3	13	13	33	16
Not take any recommendation	0	3	10	17	8

Although we observe from figure 5.2 that users tended to take at least one recommendation when they were either neutral or agree that the recommendation list contains movies they think they would enjoy watching[2], a fisher exact test shows that users' taking recommendations is independent from their perceived enjoyment (p.value = 0.45). This insignificant statistical test suggests that the self-reported enjoyment does not relate to the user subsequent consumptions. In other words, inferring the user subsequent consumption based on their self-reported enjoyment can lead to bias in understanding their satisfaction.

Perceived personalization. Similarly, we look at the answers of 116 users for the 2^{nd} question presented in table 5.1. In table 5.3, we present the number of users per answering scale and whether they took at least one recommendation or not. In figure 5.3, we visualize the percentage of users who took recommendations per answering scale.

We observe the same pattern described above. Although it seems to have a pattern which suggests that the more users agreed the recommendation lists were personalized to them, the more likely they would take at least recommendation

[2] Less than 20% of users who did not enjoy the presented recommendations actually took at least one recommendation (and 0% of users who strongly did not enjoy the recommendations took at least one recommendation). On the other hand, more than 30% of users answered that they enjoyed recommendations (either *Agree* or *Strongly Agree*) did take at least one recommendation. Users who enjoyed or did not enjoy the recommendations were those who were more likely to try out at least one recommendation - with 43% of them took at least one recommendation.

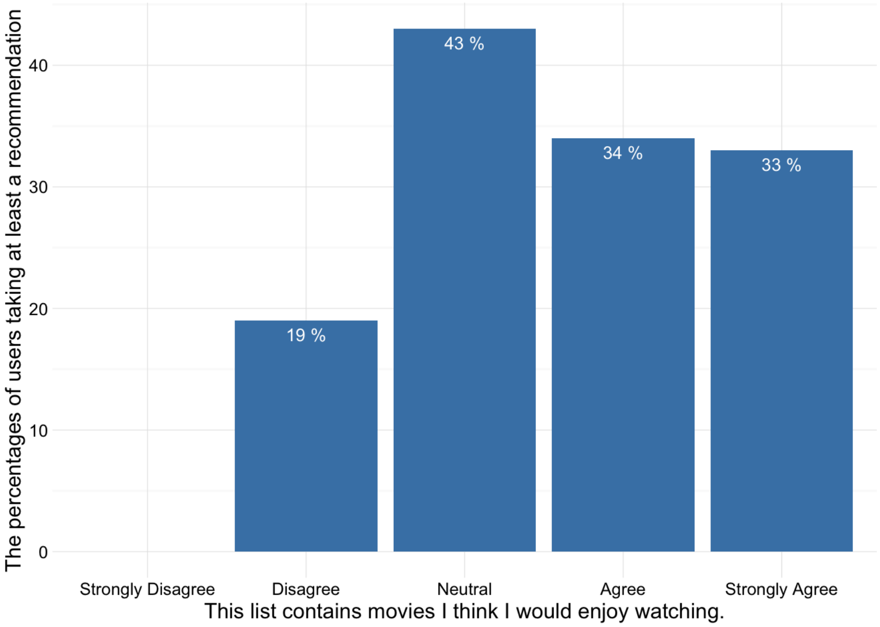


Figure 5.2: The percentage of users taking at least one recommendation from a recommendation list with different levels of perceived enjoyments.

from the list, a fisher exact statistical test suggests that there is no dependency between the actual user consumption of recommendations and the user perceived personalization (p.value = 0.66).

Table 5.3: The number of users per answer for perceived personalization and whether they took at least a recommendation from the list.

	Strongly Disagree	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
Take at least one recommendation	4	13	24	27	10
Not take any recommendation	0	5	12	17	4

Perceived bestness. We are also interested in knowing whether user actual recommendation consumption depends on how easily they thought of other movies they would rather watch than the recommendations. Thus, we look at the answers of 116 users for the 3^{rd} question, count the numbers of users per answer per whether they took at least one recommendation. These numbers are as shown in

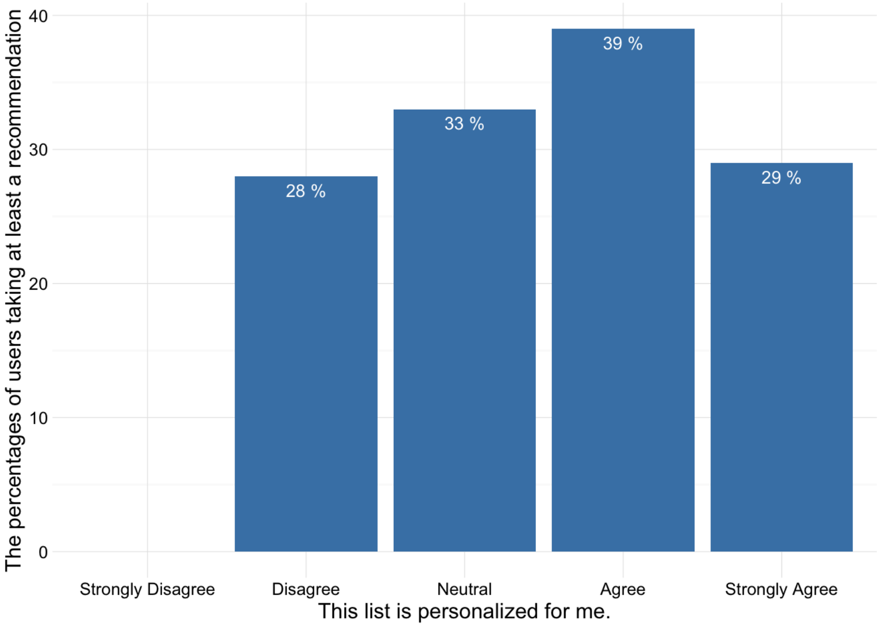


Figure 5.3: The percentages of users taking at least one recommendation from a recommendation list with different levels of perceived personalizations.

table 5.4.

Table 5.4: The number of users per answer for whether they can easily think of movies they rather watch and whether they took at least a recommendation from the list.

	Strongly Disagree	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
Take at least one recommendation	1	13	28	20	16
Not take any recommendation	1	6	8	14	9

We observe that users who agreed or disagreed that they could easily think of movies they rather watch than those on the list were more likely to take at least one recommendations than those who took a neutral stand on the question. This suggests that we do not have any evidence that the user recommendation consumption depends on how users easily think of other movies. A fisher exact test also confirms the null hypothesis (p.value = 0.44).

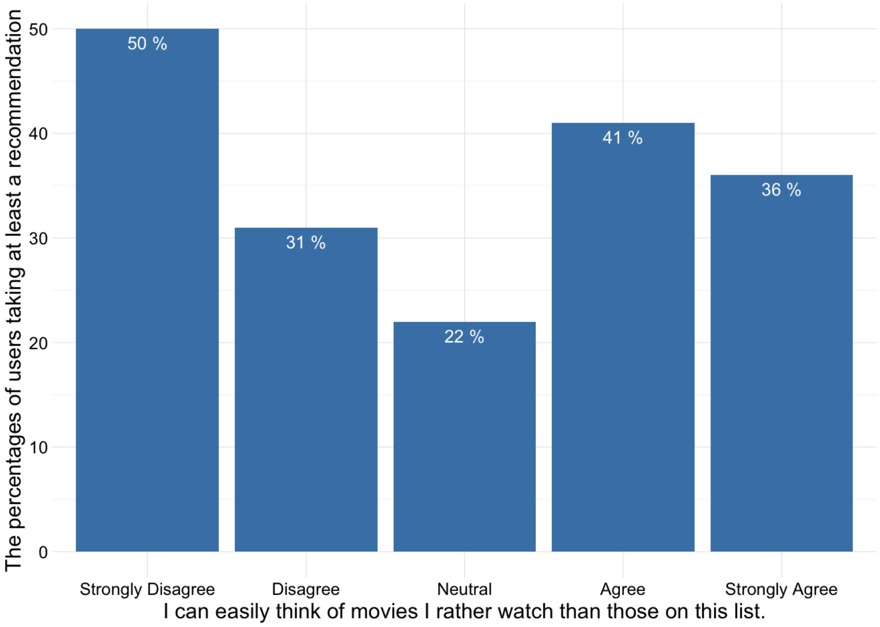


Figure 5.4: The percentages of users taking at least one recommendation from a recommendation list with different levels of thinking of other movies.

Satisfaction with diversity.[3]. We move on to look at the user satisfaction with recommendation diversity, popularity, and serendipity. Because the 4^{th} question to evaluate user satisfaction with recommendation diversity has the satisfactory option at the middle of the answering scale, we classify users who answered the question *'Just right for me'* as satisfied users. For other users, we classify as non-satisfied users. Of 116 users, we count the number of users per satisfactory levels and whether they took at least one recommendation as shown in contingency table 5.5.

We observe that only 28% of non-satisfied users took at least one recommendation from the presented recommendations. This percentage jumps to 36% for satisfied users with recommendation diversity (the two left-most columns in figure

[3] Because the correlations between users' self-reported satisfaction are small (diversity vs. popularity: -0.19, diversity vs. serendipity: 0.15, popularity vs. serendipity: -0.20), we analyze the user satisfaction with diversity, popularity, and serendipity independently.

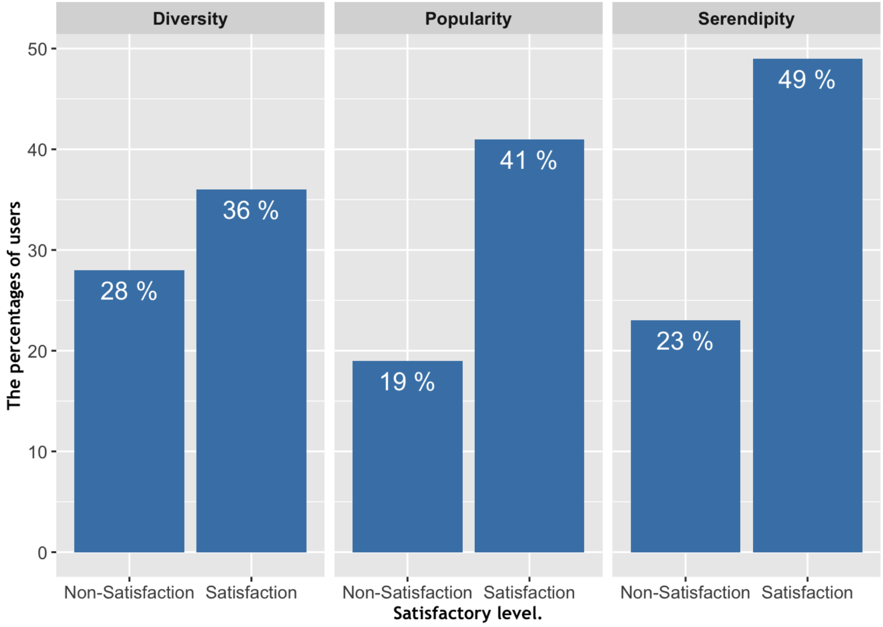


Figure 5.5: The percentages of users taking at least one recommendation from a recommendation list when satisfied or not about the recommendation list diversity, popularity, and serendipity.

5.5. However, although the data suggests that users would take recommendations if they were satisfied with the recommendation diversity, a fisher exact test suggests this dependency is not statistically significant (p.value = 0.43).

Satisfaction with popularity. Similar to recommendation diversity, we classify users who answered *'Just right for me'* for the 5^{th} question in table 5.1 as satisfied users. For other users, we classify as non-satisfied users. Table 5.6 shows the counts of users per satisfactory levels and whether they took at least one recommendation. From table 5.6 and figure 5.5, we observe that only 19% of users who did not satisfy with the recommendation popularity consumed at least one recommendation. However, this percentage jumps to 41% for those who were satisfied with recommendation popularity. A fisher exact test shows that indeed user actual recommendation consumption and their satisfaction with popularity are dependable (p.value = 0.01).

Table 5.5: The contingency table of the user satisfaction with recommendation diversity and their recommendation consumption.

	Not satisfied with recommendation diversity	**Satisfied with recommendation diversity**
Not take a recommendation	38	40
Take at least one recommendation	15	23

Table 5.6: The contingency table of the user satisfaction with recommendation popularity and their recommendation consumption.

	Not satisfied with recommendation popularity	**Satisfied with recommendation popularity**
Not take a recommendation	35	43
Take at least one recommendation	8	30

Satisfaction with serendipity. Like diversity and popularity, we classify users who answered *'Just right for me'* for the 6^{th} question in table 5.1 as satisfied users, otherwise non-satisfied users. Table 5.7 shows the counts of users per satisfactory levels and whether they took at least one recommendation. From table 5.7 and figure 5.5, we see that only 23% of users who did not satisfy with the recommendation serendipity consumed at least one recommendation. However, this percentage jumps to 49% for those who were satisfied with recommendation serendipity. A fisher exact test on the counts presented in table 5.7 shows that indeed the user recommendation consumption and their satisfaction with serendipity are dependable (p.value = 0.00).

5.3.2 Recommendation properties and the inferred user recommendation consumption

For the 1^{st} research question we find that only the self-reported user satisfaction with recommendation popularity and recommendation serendipity are good indicators if users would take at least one recommendation from the list. We move on to answer our 2^{nd} research question.

Table 5.7: The contingency table of the user satisfaction with recommendation serendipity and their recommendation consumption.

	Not satisfied with recommendation serendipity	Satisfied with recommendation serendipity
Not take a recommendation	56	22
Take at least one recommendation	17	21

RQ2: Whether and how the three properties of recommendations and recommendation lists relate to the inferred user recommendation consumption?

To answer this question, we look at the recommendation diversity, recommendation popularity, and recommendation serendipity.

Recommendation diversity, popularity, and serendipity. For diversity, we compute a diversity score per recommendation list based on equation 5.1. For popularity, we compute a popularity score per recommendation list based on equation 5.2. For serendipity, we compute a serendipity score per recommendation list based on equation 5.3.

To answer our research question, we run the below logistic regression:

$$\begin{aligned} taken.recommendation \sim diversity.score + \\ popularity.score + serendipity.score + (1|userId) \end{aligned} \tag{5.4}$$

where *taken.recommendation* indicates whether users took at least one recommendation. In this model, users are treated as random factors. We have three scores in one logistic regression model because we want to control popularity levels and serendipity levels of recommendation lists while testing for diversity levels. Likewise for popularity and serendipity levels. Furthermore, the three scores are not correlated[4].

Table 5.8 shows the results of the logistic regression. Based on the coefficients in the table, we observe that the more diverse or the more popular the recommendations are, the more likely users would take at least one recommendation.

[4] The correlation between serendipity and popularity is -0.35; between diversity and popularity, is -0.55; between diversity and serendipity is 0.44.

Table 5.8: The logit coefficients for recommendation diversity, popularity and serendipity with respect to the probability of users taking at least one recommendation.

	Estimated	**Std. Error**	**Z value**	**p.values**
Intercept	1.73	2.01	0.86	0.39
Diversity	6.41	4.40	1.46	0.15
Popularity	0.03	0.06	0.59	0.56
Serendipity	-12.22	5.02	-2.43	0.02

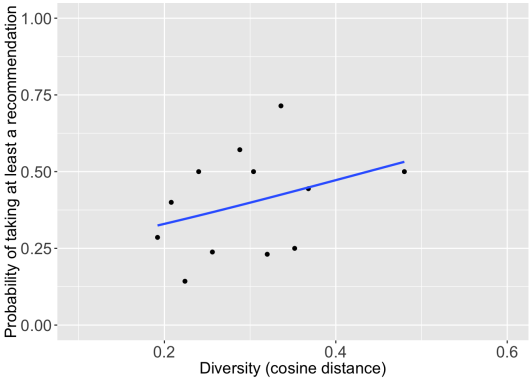


Figure 5.6: The probability of users taking at least one recommendation from a recommendation list at different amounts of recommendation diversity.

We also observe that the more serendipitous the recommendations are, the less likely users would take at least one recommendation. In fact, these trends can be seen in figures 5.6, 5.7, and 5.8. Interestingly, only the finding about serendipity is statistically significant. The significant coefficients about serendipity suggests that for every 10% increase in serendipity, the odd changes in the probability of users taking a recommendation is 0.29

5.4 Discussion

We set out to understand the relations between the self-reported user satisfaction and the subsequent user consumption in a recommender system. We also investigate the relations between recommendation properties and the subsequent

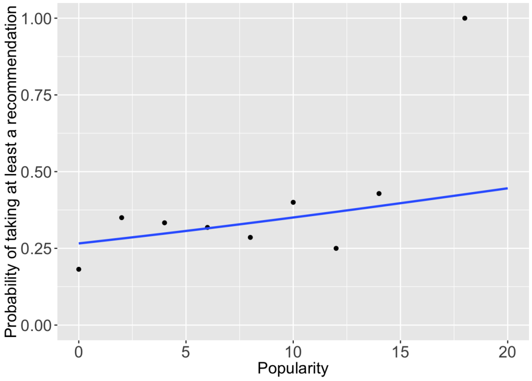


Figure 5.7: The probability of users taking at least one recommendation from a recommendation list at different amounts of recommendation popularity.

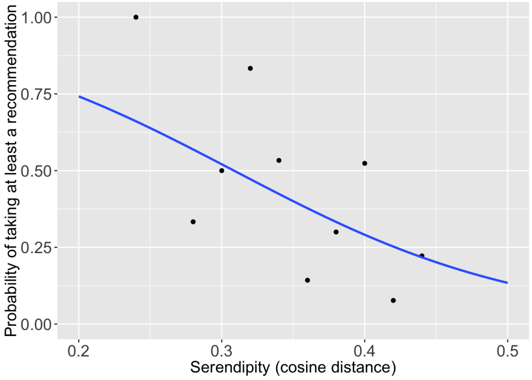


Figure 5.8: The probability of users taking at least one recommendation from a recommendation list at different amounts of recommendation serendipity.

user consumption. To that end, we look at the following factors: 1) the user self-reported satisfaction explicitly measured by survey questions; 2) the diversity, popularity, and serendipity of recommendations and recommendation lists. We then analyze how these factors relate with the inferred user recommendation consumption.

We found that self-reported user satisfaction with popularity and serendipity related to subsequent user recommendation consumption. Users who were satisfied with recommendation popularity and serendipity, as measured via our survey questionnaire, would be more likely to take recommendations than those who were not. Subsequently, we found the same patterns while investigating the relations between the three recommendation properties (popularity and serendipity) and the subsequent user recommendation consumption. These patterns strengthen the aforementioned results. We also found evidence suggesting that users who were satisfied with recommendation diversity would also be more likely to take recommendations. However, this evidence is not statistically significant. This insignificant result begs the question whether most users want to see diverse lists yet, few of them want to commit to recommendations that are too different for their tastes. Investigating this insignificant result is out of scope for this paper, and we will look into this in our future work.

Interestingly, we do not have any statistical evidence suggesting that when users expressed their enjoyment for a set of recommendations, they would take at least one recommendation from the set. These are surprising results since we hypothesized that the more users perceived that they would enjoy watching these recommendations, the more likely they would take recommendations. Likewise for the perceived personalization. Surprisingly, despite documented evidence in the literature that diversity correlates with user satisfaction, we found no statistically significant relations between recommendation diversity with user recommendation consumption. We hypothesize that in movie-recommendation domain, popularity and serendipity are more important factors than diversity. As mentioned above, users might want to see diverse lists, but not to actually commit to consuming them.

To summarize, in answering our research questions, we make several contributions. First, we provide quantitative evidence suggesting that researchers and practitioners should prioritize to evaluate user satisfaction with recommendations at the time they carry out user studies with questionnaires. Second, we find evidence that combining longitudinal logged data study with questionnaire-based user study would give a better holistic view of user satisfaction. Third, we find evidence that of the three recommendation properties (diversity, popularity, and serendipity), serendipity has the strongest influences on user satisfaction in movie-recommender systems. Thus, we recommend researchers and practitioners to focus more on serendipity than on popularity and diversity for recommendations.

Besides, our study has several implications. First, we provide a quantitative evidence that user satisfaction with recommendation popularity and serendipity are the most important due to its connections to the subsequent user recommendation consumption. Thus, recommender system designers and researchers should focus on these factors when implementing recommender systems, and when studying user satisfaction with recommender systems. Second, we find that other factors such as perceived enjoyment and personalization do not necessarily[5] correlate with the user recommendation consumption. Thus, we think that studying user satisfaction with surveys alone might not be sufficient to understand how users are satisfied with the recommendations. Because of the uncontrollable external factors, users might not be able to express their desires accurately when answering the survey questions, causing potential misleading interpretations about their satisfaction. We recommend combining survey questions with a longitudinal study to limit any undesired bias.

Last but not least, our study is not without limitations. First, our study focuses on one domain of recommendations - movie recommendations. Thus, we hesitate to generalize our conclusions to different domains of a recommender system. Second, there is a temporal gap between when users take the survey evaluating their satisfaction with recommendations, and when they took recommendations. Users' impression of the recommendation list may fade away. There may be other possible measurements of the user post-recommendation consumption that can attenuate users' fading memory on their satisfaction of recommendation list at the time they consume the movie. We leave the future work to look into these limitations.

[5] We use the word *necessarily* because our insignificant results perhaps are due to a lack of statistical power

Chapter 6

Rating Interfaces[1]

Herlocker et al. coined the term *magic barrier* to refer to the supposition that there may be a lower bound on the minimum error that can be achieved by any ratings-based recommender system [87]. The idea is that user ratings include some level of noise, as evidenced by inconsistencies observed when asking users to rerate previously rated items [52]. One interpretation of the magic barrier is that it is unlikely that new algorithms can produce large decreases in recommender error. Another interpretation, and the one that motivates this research, is that if we can understand the cognitive processes that drive the magic barrier then perhaps we can create new ways for users to rate items, reducing the minimum achievable error in our systems.

Psychological research on preference construction and expression provides insight into how users go about rating items. This research suggests that most users do not form a rating at the time of consumption and store that rating in their memory [62]. Rather, they store a complex set of thoughts, feelings, perspectives

[1] This chapter is based on the work of Nguyen et al. [75] published in the 7^{th} ACM conference on Recommender systems. The author of this dissertation led this research project, developed algorithms for interfaces and back-end infrastructures to store all the necessary data, prepared data for the experiment, helped in analyzing the data, and writing the paper. Daniel Kluver is also the first co-author of this paper and the first author of the paper [57] which provides one of the frameworks to analyze the results in this paper.

and insights, and map those complex stored values into a rating when they are asked to provide one. This process is labeled preference construction [63]. There are many ways this process can go poorly, potentially increasing the noise, and raising the magic barrier.

In this work we address two likely sources of errors in the rating process. The first is that users may not clearly remember items, especially if the users experience these items a long time prior to rating them. The second is that users may struggle to accurately and consistently map their constructed preferences to the rating scale.

It seems plausible that modifying the rating interface may mitigate these problems. To that end, we study four interfaces inspired by these two key methods of support: one that is a baseline, one to help users *remember movies*, one to help users *understand their personal rating scale*, and one combined interface to help *with both* simultaneously.

We explore a novel set of methodologies to examine the effectiveness of the interfaces by performing a re-rating experiment involving 386 users and 38,586 ratings in MovieLens. We apply the information theoretic framework developed in our previous work [57], and the magic barrier framework developed by Said et al. [88] to measure the effect of our interfaces on the consistency of, and noise in, a user's ratings. Furthermore, we investigate the impact of these interfaces on cognitive load and user satisfaction via a user survey.

Noise in user rating is one of the main challenges in recommender systems. We argue that we can improve the quality of rating data by implementing interfaces that help users map their constructed preference to the rating scales better. With this in mind, our contributions are novel interface designs to collect user data with less noise, and a set of methodologies to evaluate the effectiveness of these interfaces.

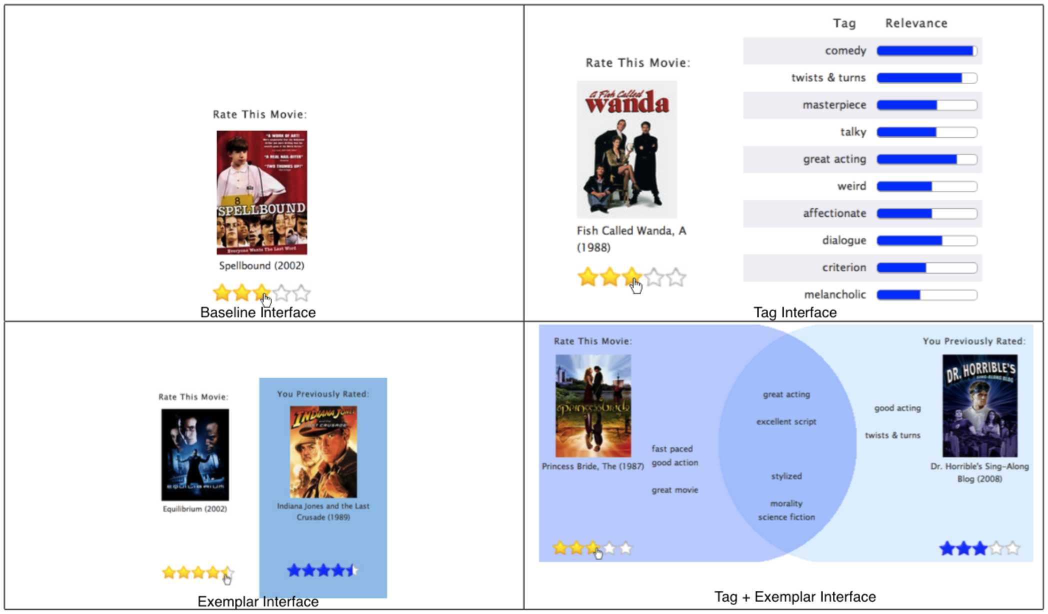


Figure 6.1: Our four interface designs to provide memory-support, rating support and joint-evaluation support.

6.1 Interface Design

To answer the research questions, we develop four interfaces: one with minimalistic support that serves as the baseline, one that shows tags, one that provides exemplars, and another that combines the previous two features (figures 6.1).

Figure 6.1 (top left) presents the baseline design allowing users to map their internal preferences into ratings. To assist our users in the rating process, we provide basic information such as posters and titles. However, posters and titles might not provide sufficient memory support. Therefore, we design the tag interface that provides personalized memory support.

6.1.1 The Tag Interface

One cause of rating noise is in the memory processes involved in recalling the item. As our memory is reconstructive, different aspects will come to mind at different times and in different situations [64], resulting in different constructed preferences. Our memories might also become less detailed and precise over time, and the importance of certain aspects might change, leading to our ratings' changes over time [89]. System designers help users form consistent and useful preferences by helping the user recall the item. They do this by presenting descriptions and attributes of the item that will help the user remember the item. We call these features 'memory support' as they help users remember the item.

For memory support to be helpful, we argue that it should be tailored to the user as much as possible. If we are able to predict which aspects are most important for a user, we can use these aspects to provide more effective personalized memory support than general background information. Therefore we compare memory support interfaces using personalized tags against a minimal support interface.

In our design process, we look for a personalizable form of memory support that easily triggers the recall process. While some users remember a movie because it is funny, others remember the movie because it is visually appealing. To provide personalized information about the movie, we use the tag genome [14] combined with users' previous ratings. The tag genome is a tag-movie relevance matrix, inferred from user-supplied tags, ratings, and external textual sources. The relevance of a tag t to a movie m determined by the tag genome is a value between 0 and 1 denoted as $rel(t, m)$. We manually remove two classes of tags from the genome prior to rendering our interfaces. The first is tags such as 'excellent' and 'great' that indicate a general quality judgement of the movie; these are redundant with ratings and do not provide any information about the movie. The second is binary metadata tags such as directors, actors, 'based on a book', etc. These tags describe properties that are objectively verifiable and either true or false; it is incorrect to infer them when they are not explicitly added to the item. Future work should build a way to incorporate such metadata tags, but for now we restrict our use of the tag genome to tags for which inference makes sense. Figure 6.1 (top right) shows the tag interface.

To personalize the interface, we select tags that best help users reconstruct their memory about the movies. Due to the limited space on the interface, only a set of 10 personalized tags from 1,574 candidates can be displayed. To address this problem, we use Nguyen et al.'s approach [90] to compute how much information about user u's preference a tag t can tell, denoted as $pMI(t, u)$. We hypothesize that the more information a tag can tell about user u's preference, the more likely it is helpful to the user. Moreover, a useful tag must be sufficiently relevant to the movie. Therefore, we require that any selected tag must have a minimum relevance of 0.25.

Similar tags, such as 'political' and 'world politics', diminish the effectiveness of memory support if they are displayed together. To address this challenge, we place a similarity constraint on the tag selection algorithm. Our prototyping suggests that mutual information between tags works well. Formally, let $P(t_1)$ be the probability that tag t_1 is applied to a random movie, $P(t_1, t_2)$ be the joint-probability that tags t_1 and t_2 are both applied to a random movie. Hence, the tag similarity, denoted as $dTagsim(t_1, t_2)$, is defined as $dTagSim(t_1, t_2) = \frac{I(t_1,t_2)}{H(t_2)}$, where $I(t_1, t_2)$ is the mutual information of the two tags, and $H(t_2)$ is the entropy of tag t_2. Note that dTagSim, or directed tag similarity, is asymmetric, i.e. $dTagSim(t_1, t_2) \neq dTagSim(t_2, t_1)$. After several trials, we find that 1% (0.01)to be a reasonable cut-off.

With these concepts laid out, our tag selection approach is a constrained optimization problem described as follow:

$$\begin{aligned} \underset{T^\star}{\text{argmax}} & \sum_{t \in T^\star} rel(t, i) \times pMI(t, u) \\ \text{subject to } & \mid T^\star \mid = 10 \\ & dTagSim(t_i, t_j) < 1\% \ \ \forall\, t_i, t_j \in T^\star,\ i \neq j \\ & rel(t, i) > 0.25 \ \ \forall\, t \in T^\star \end{aligned}$$

Since finding sets of perfect personalized tags is computationally expensive, we implement a greedy algorithm to find an approximate solution. Starting with an empty set of tags, the algorithm iteratively selects the tag with the highest $rel(t, i) \times pMI(t, u)$ such that when added to the result the constraints are still satisfied.

6.1.2 The Exemplar Interface

Noise in ratings can also be caused by difficulty in mapping a formed preference to a rating. Past research shows strong impact of anchors [91] and task characteristics [92, 93] on rating responses, when a user lacks a good internal representation of the rating scale. While it is intuitive that three stars means less preference than four stars, it can be difficult for a user to decide (in an absolute sense) whether to rate an item three or four stars. System designers help users represent their preference by making the rating scale more intuitive. A common way to do this is by providing community guidelines mapping each rating to a simple description such as: (4 stars = 'I liked it'). We call these features 'rating support' as they help convert a preference to a formal rating. However, simple labels won't be sufficient to solve the mapping problem. Tailored rating support is needed to help the user express the evaluation in a consistent way.

Prior work on rating interface design [94] has suggested using the users' past ratings to help them make future ratings. In particular they suggest an interface in which rating options are annotated with past ratings will help people make ratings that are ordered consistently. Indeed, research on preference elicitation has investigated the differences between comparing items on their own (single evaluation) and comparing items against each other (joint evaluation) [92]. When items are evaluated in isolation, the evaluation is absolute and will be predominantly based on aspects that are easy to evaluate. Using joint evaluation allows users to be more sensitive to dimensions on which items vary. We will support joint evaluation in our interface by showing exemplars for each point on the rating scale. These exemplars will be other similar items rated in the past that serve as personalized anchors on the scale, e.g.: ‘you previously rated this item 3 stars’. Research on comparative evaluation [95] suggests these exemplars should be similar to serve as good anchors otherwise it might be hard to make any comparison (or even contrast effects can occur). We will compare this interface with exemplars (supporting joint evaluation) to an interface without exemplars (single evaluation).

The exemplar interface improves upon the baseline by showing exemplars. An exemplar for a rating-movie pair is a movie that the user previously gave the same rating. Figure 6.1 (bottom left) shows the design of the exemplar interface. To support the joint evaluation, we provide anchoring points by annotating rating options with the exemplars. Each exemplar at a specific rating option is the most similar movie to the movie to rate based on the cosine similarity of the movie rating vectors. At any rating option on the rating scale, if the user has not rated any movie, a blank image is shown.

6.1.3 The Tag + Exemplar Interface

To investigate the effect of combining ‘memory support’ and ‘rating support’, we develop the fourth interface as seen in figure 6.1 (bottom right). In this interface, users make use of personalized tags to compare their rated movies to the movie to rate. When users hover over a star, an exemplar is shown. Each tag then moves either flush left, or flush right to denote which movie the tag is more relevant to (the movie to rate or the exemplar respectively). The tag will move to the center if the tag is equally relevant to both movies (relevance values within 0.1 of each other). To help the users understand the movement of the tags a Venn diagram is shown in the background of the interface. This diagram reinforces the idea that tags on one side or the other best describe the associated movie, and tags in the center describe both movies equally well.

The tag + exemplar interface uses the same exemplar selection strategy as the exemplar interface. Selecting tags, however, is more complicated in this interface.

We want tags that not only are relevant to the user and movie, but also may help the user compare between movies. We measure this in two ways. First, for each tag t we measure $\sigma(t)$ the standard deviation of the relevance values of each chosen exemplar to that tag. If a tag's relevance varies greatly between the exemplars then it can help serve to differentiate them.

Secondly, we measure how well a tag splits between being more relevant to the movie to rate, or the exemplars. If a tag is always more relevant to the movie to rate, or always more relevant to the exemplars, then the tag does not help the user compare the different rating choices. To measure this we define $Z(i,t) = \frac{\max(M_t, L_t)}{\min(M_t, L_t)}$, where M_t is the number of exemplars to which tag t has higher relevance than to the movies to rate, and L_t is the number of exemplars to which tag t has lower relevances than to the movies to rate. When the tag would appear with the exemplar as often as with the movie to rate then $Z(i,t) = 1$, as the balance becomes less equal $Z(i,t)$ becomes larger.

Our algorithm for this tag + exemplar interface is also a constraint optimization problem described as follow:

$$\begin{aligned} \underset{T^\star}{\text{argmax}} \; & \sum_{t \in T^\star} rel(t,i) \times pMI(t,u) \times \sigma(t) \div Z(i,t) \\ \text{subject to} \; & \mid T^\star \mid = 10 \\ & dTagSim(t_i, t_j) < 1\% \;\; \forall \, t_i, t_j, \in T^\star, \; i \neq j \end{aligned}$$

Since finding sets of perfect personalized tags is computationally expensive, we implement a greedy algorithm to find an approximate solution. Starting with an empty set of tags, the algorithm iteratively selects the tag with the highest $rel(t,i) \times pMI(t,u) \times \sigma(t) \div Z(i,t)$ such that when added to the result the constraints are still satisfied.

6.2 Experimental Design

To compare the our interfaces, we ask users to rate 50 movies, and re-rate these movies after at least two weeks. Users are asked to answer 22 survey questions after finishing the second round. To help users understand the interfaces, we provide a tutorial for each interface explaining the interface and walking users through the interface step by step. Our analyses are based on techniques described as below.

6.2.1 Efficacy Metrics

RMSE. One of the hopes in improving rating quality is to reduce the magic barrier of a recommendation system. One easy to measure consequence of reducing the magic barrier is to reduce the RMSE of a recommendation system. To measure this effect we estimate the minimum RMSE possible over the ratings using the technique introduced by Said et al. [88]. Said et al. argue that the average variance of ratings (measured at each user item pair) gives an estimate of the minimum possible RMSE over these ratings. If any of our interfaces lower the magic barrier we should see a lower minimum RMSE. To confirm our findings with this method we also build simple item-item recommenders for each interface and compare the RMSE on these interfaces directly.

Although the RMSE quality metrics can suggest that we have lowered the magic barrier, these metrics do not capture the whole picture of rating quality. For example, RMSE can trivially be reduced by influencing users to use a smaller range of the rating scale, which would lead to overall less informative ratings. An absolute measure of the information and noise contained in ratings could be a more useful tool in evaluating the efficacy of the interfaces. Therefore, in the next two sub-sections, we propose approaches to measuring the amount of noise and information about user preferences in ratings collected from each interface.

Preference Bits per Rating. In prior work [57] we showed how to measure the preference bits of ratings from an interface. This framework can be used to estimate the total amount of information ratings contain about user preferences. An increase in preference bits from any interface would indicate that ratings on that interface tell us more about users' preferences.

We measure preference bits by gathering two ratings for each user-item pair and computing the mutual information between the two sets of ratings. This measures how much the first round of ratings reduces our uncertainty about ratings in the second round, providing a lower bound for the information that one rating contains about the preference that caused it. If ratings on one interface give us more information about user preferences than ratings on another interface, we say that the interface is more effective at eliciting preference.

Rating Noise. It is possible for ratings from different interfaces to contain the same amount of information with different amounts of noise. Therefore, even if our interfaces yield the same amount of preference information, we may still prefer one with less noise. Information theory gives us a clear way of measuring the amount of noise contained in our ratings.

Let π and R be random variables representing the user's true preference for, and rating on, an item respectively. The total randomness in ratings is measured by the entropy of the ratings $H(R)$. Relative to any other random variable X, the entropy of R can be split into two parts, the mutual information $I(R;X)$ measuring the amount of information R contains about X, and the conditional entropy of R given X, $H(R|X)$. Taking this decomposition with respect to the user's true preference π we get the following identity.

$$H(R) = I(R;\pi) + H(R|\pi) \tag{6.1}$$

Based on equation 6.1, we will measure the amount of noise contained in ratings by the conditional entropy of a rating given the user's preference, $H(R|\pi)$. Because true preference π cannot be measured we will have to estimate the true rating noise. In our previous work [57] we showed that we can estimate the true preference bits using two ratings for the same user-item pair taken at different times R_1, R_2,. We prove a similar result about the rating noise. Formally in previous work we showed:

$$I(R_1; R_2) < I(R; \pi) \tag{6.2}$$

Using equations 6.1 and 6.2 we have the following:

$$H(R|\pi) = H(R) - I(R; \pi) < H(R_1) - I(R_1; R_2) = H(R_1|R_2)$$

The conditional entropy between two ratings is therefore an upper limit to the true rating noise. We will use this metric as an estimate of the true rating noise of our interfaces. Because recommendation algorithms are likely sensitive to noise in the ratings, we hypothesize that we can learn better from ratings with less total noise.

6.2.2 User Experience Metrics.

Objective User Experience (Cognitive Load). Harper et al. [96] investigate motivations behind user rating behaviors in a recommender system domain, and find that users perceive rating-time costs when they provided ratings. Harper et al. also point out that the users keep providing ratings when they perceive that the benefits outweigh the costs they pay. Sparling et al. [58] investigate further to estimate the mental cost and benefits on different rating scales. Since accurately measuring user mental costs is hard, we follow Sparling et al.'s approach to use rating time to estimate cognitive load. We use this metric, as well the user experience metric below, to evaluate the trade-offs between getting more information about user preference and overloading users.

Subjective User Experience (Self-report). To learn how users perceive the usefulness and difficulty of the four interfaces, we conclude our experiment with a survey. We then follow Knijnenburg et al.'s [97] evaluation framework for user experience. The framework models how objective system aspects such as the user interface influence users' subjective perceptions and experience, controlling for other factors such as situational characteristics and personal characteristics. Particularly, we measure how usefulness of the system (experience) is influenced by the perceived difficulty of the system (subjective system aspect), and how each of these are affected differently by the 4 different interfaces, controlling for the self-reported expertise as a personal characteristic. The survey consists of 22 7-likert-scale statements, ranging from 'Strongly disagree' to 'Strongly agree', that query the theoretical constructs of interest: usefulness of the interface, difficulty of the interface and self-reported movie expertise. Some questions are inverted to check users responses for coherence and lack of effort. Participants have to answer all questions. They can change their answers before submission. Participants can provide extra feedback in a text-box. We also ask (but not require) participants' age and gender (refer to table 6.2).

6.2.3 Design

We conduct our study with users of MovieLens,[2] an online collaborative filtering-based movie recommender system. In order to make our algorithms work on these interfaces, we select only users who have rated at least 100 movies that have both tag genome and Netflix title data available. We also require that the 100 ratings be provided since MovieLens changed to its current half-star rating scales [3]. Finally, we require that users have logged in at least once during the last four years. After preparing data for the experiment (2013-03-01), we randomly invited 3829 users via email invitations. The users were randomly assigned to one of the four interfaces.

Our participants are asked to rate 50 movies in the first round, and re-rate these 50 movies in the second round. To simulate the real rating experience, the participants are asked to rate the most recent 50 movies that they rated prior to the study. To avoid users recalling recently rated movies, we wait at least 14 days

[2] http://www.movielens.org/

[3] We estimate that MovieLens switched to a five-rating-star with half-star increment on 2003-02-12

after data preparation before inviting users to participate in our study (round 1). This guarantees that at least two weeks have passed since the user last rated the selected movies in MovieLens. For the same reason, we also wait 14 days before starting round 2. After finishing the second round, participants are asked to complete a survey asking their experience with the interfaces as well as the rating process. For each 20 participants, a randomly selected one received a $50 Amazon Gift card.

6.3 Results

Interface	Minimum RMSE	RMSE (round 1)	RMSE (round 2)	Preference Bits per Rating [4]	Rating Noise	Mean Rating Time	Median Rating Time
Baseline	0.232	1.13	1.12	1.26	1.71	4.26	3.71
Exemplar	0.209	0.95	0.95	1.22	1.62	5.53	4.60
Tag	0.237	1.03	1.10	1.22	1.72	4.49	4.07
Tag + Exemplar	0.227	1.04	1.01	1.25	1.70	6.70	5.60

Table 6.1: Our quantitative results for four interfaces

Our experiment ran from March 21st 2013 until April 25th 2013, collecting 38,586 [5] ratings from 386 users, of which 103 users were assigned the baseline interface, 100 to the exemplar interface, 98 to the tags interface, and 85 to the tags + exemplars interface. 352 provided information about age (mean: 36, min: 18, max: 81, σ : 11). 304 were male, 72 female, and 10 did not report. Our analyses don't reveal significant differences between the two genders.

6.3.1 Efficacy Metrics

RMSE. We follow the procedure described in [88] to estimate the minimum RMSE for each interface. The estimated minimum RMSE for our interfaces are given in table 6.1. To test these differences for statistical significance we compute an estimated minimum RMSE for each user. Using an ANOVA on the per user

[5] Due to network issues, we had 14 incomplete ratings which were removed from all of the analyses.

minimum RMSE estimates we find marginal significance that the average per-user minimum RMSE differs between interfaces ($p = 0.0574$). Using a TukeyHSD we find marginal evidence that the exemplar interface has a lower average minimum RMSE than the tag and baseline interfaces (adjusted $p = 0.0689$, 0.098 respectively).

To control for possible error in the ANOVA due to violations of the normality assumption we confirm these results with a pairwise Wilcoxon rank sum test over the per user minimum RMSE using the Holm method to correct for multiple comparisons. We find that the minimum RMSE is significantly different between the exemplar interface and baseline($p < 0.005$), however, we find only minimal support for the conclusion that the exemplar interface differs significantly from the tag interface ($p = 0.208$). Therefore we conservatively conclude that the only significant pairwise difference in our minimum RMSE estimates is between exemplar and baseline.

To compare these results against actual predictive accuracy we use the LensKit toolkit [98] to train and evaluate an item-item collaborative filtering recommender over the ratings from the four interfaces. We use 10-fold cross validation with a 10 item holdout set for each test user. Table 6.1 shows the per-user RMSE of each condition for each round. The exemplar interface has the lowest RMSE in both rounds. The measured RMSEs are significantly different ($p < 0.001$ using ANOVA on the user RMSEs); the exemplar had lower RMSE than the baseline in round 1 ($p < 0.001$ using Tukey HSD post-hoc test), and from both the baseline and tag interfaces differ significantly in round 2 ($p < 0.05$).

Preference Bits per Rating. We estimate preference bits using the technique described in [57]. The results are summarized in table 6.1.

To our knowledge significance testing on information theoretic values is not a widely studied field. Therefore, to test for significance in our measured mutual information differences we use a resampling based strategy to test each pairwise difference. Under the null hypothesis ratings from any two interfaces are drawn from the same distribution (and therefore have the same true mutual information). To test for differences between any two conditions we randomly permute the assignments of users to conditions. It is important to permute interface labels over users rather than ratings to control for outlier users who rate with significantly more or less information or noise than the average user. For each permutation we compute the difference in means of measured mutual information between the two conditions. We reject our null hypothesis if fewer than 5% of our random trials have differences of means as large or larger than observed. We perform this test using 10000 permutations and find that no pairwise difference in mutual information is statistically significant. Therefore we do not have sufficient evidence to conclude that our interface modifications had a significant effect on the amount of preference information extracted.

Rating Noise. The preference noise of an interface is estimated by the conditional entropy of the first rating given the second rating in a rerating dataset. The conditional entropy estimates for each interface can be seen in table 6.1. The rating noise metric generally agrees with the RMSE based metrics, especially in measuring the exemplar interface as having the least noise. To test these measurements for differences we use the same resampling strategy discussed above. Using 10000 random permutations we find that no pairwise difference in conditional entropies is significant. Therefore we do not have sufficient evidence to conclude that our interface modifications had a significant effect on rating noise.

It is interesting to note that while we are not able to conclude that our interfaces have an effect on rating noise, we are still able to conclude effects on the RMSE of the ratings. This result is contrary to our belief that noise in user ratings is a primary cause of the magic barrier. We expect this finding is largely due to differences in power between our statistical tests for rating noise, and the RMSE based metrics. Because statistics on information theoretic measures is a not-well explored field of statistics, it is possible that analysis with a more powerful technique would allow us to find significance in our rating noise metric that match the results found in the RMSE section. Nonetheless we feel that the rating noise metric is a better metric for future use because it directly measures the noise in ratings, rather than measuring a consequence of this noise.

6.3.2 User Experience Metrics

Objective User Experience (Cognitive Load). Cognitive load is estimated by measuring the time it takes users to rate a movie. We assume that it will take users time to process the information when using interfaces. A good interface could decrease the cognitive load and time required.

Before analyzing the rating times, we first exclude long rating times that occur because users are distracted during rating or take a break. Such long times introduce bias in our analysis since these rating times do not reflect real users' cognitive loads. We exclude the top 1% of the data points for each interface, resulting in cut-off points of 41 seconds for the baseline interface, 43 seconds for the tag interface, 55 seconds for the exemplar interface, and 79 seconds for the tag + exemplar interface.

As users made 50 ratings in each round, we take the average rating time per round as an estimated measure of cognitive load. Table 6.1 shows the statistics of rating time per interface aggregated over both rounds. Our ANOVA analyses suggest that the required cognitive load for the baseline and tags interfaces be the smallest among the four (p < 0.01). The difference between the two interfaces is not significant ($p = 0.29$). This suggests that users do not take advantage of the memory support provided by the tags interface. Participants spent the most time on the tag + exemplar interface, followed by the exemplar interface (The difference is significant, p < 0.0001). All other pairwise comparison are also significant (p < 0.0001). This suggests that a) users utilize the supports provided by the two exemplar empowered interfaces[6], or b) users might find the tag and

[6] we also perform multilevel linear regression analysis with random intercepts that also take into account the 50 repeated observations per round, with similar results. We also apply separate ANOVA analyses for two rounds, and observe that in the second round, there is a marginal

exemplar interface were confused, thus they spent more time on it.

Subjective User Experience (Self-report). Our questionnaire was designed to measure user experience of usefulness and difficulty of the interface and participant's expertise. The items in the questionnaires were submitted to a confirmatory factor analysis (CFA). The CFA used ordinal dependent variables and a weighted least squares estimator, estimating 3 factors. Items with low factor loadings, high cross-loadings, or high residual correlations were removed from the analysis. Factor loadings of included items are shown in table 6.2, as well as Cronbach's alpha and average variance extracted (AVE) for each factor. The values of AVE and Cronbach's alpha are good, indicating convergent validity. The square roots of the AVEs are higher than any of the factor correlations, indicating good discriminant validity.

Considered Aspects	**Items**	**Factor Loading**
Usefulness Alpha: 0.83 AVE: 0.546	MovieLens should use this interface	0.89
	The interface helped me think carefully about how much I liked the movie	0.70
	The interface helped me express my ratings consistently	0.66
	The interface did not make rating easier compared to the usual MovieLens rating interface	-0.71
	Rating movies with this interface was fun	0.71
	The interface helped me to remember what I like and dislike about movies	
	I found the interface helpful in relating a movie to other movies I have seen	
	The interface did not motivate me to think carefully about my ratings	
	I know of other sites with rating interfaces I would rather use than this one	
	I would recommend this interface to other users	
Difficulty Alpha: 0.84 AVE: 0.71	I found the interface confusing	0.86
	The interface was easy to understand	-0.86
	I found the interface difficult to use	-0.82
	Rating with this interface was fast	
	Rating with this interface was easy	
Movie Expertise Alpha: 0.68 AVE: 0.55	I'm a movie lover	0.75
	Compared to people I know, I read a lot about movies	0.85
	Compared to people I know, I'm not an expert on movies	-0.61
	I mostly like popular movies	
	I get very excited about some upcoming movies	
	I know why I like the movies I like	
	I often like to compare a new movie to others I've seen	

Table 6.2: Items presented in the questionnaires. Items without a factor loading were excluded from the analysis.

The subjective constructs from the CFA (table 6.2) were organized into a path model using a confirmatory structural equation modeling (SEM) approach with ordinal dependent variables and a weighted least squares estimator. In the

significant difference between the two exemplar empowered interfaces p = 0.068) resulting model, the subjective constructs are structurally related to each other and to the four interface types. In the final model, the expertise factor did not relate to any other variable, and was therefore removed from the analysis. The final model had a good model fit [7] ($\chi^2(52) = 3408$, $p < 0.001$, CFI = 0.979, TLI = 0.971, RMSEA = .070, 90% CI: [.054, .085]). Figure 6.2 displays the effects found in this model. Factor scores in the final model are standardized; the numbers on the arrows ($A \rightarrow B$) denote the estimated mean difference in B, measured in standard deviations, between participants that differ one standard deviation in A. The number in parentheses denotes the standard error of this estimate, and the p-value below these two numbers denotes the statistical significance of the effect. As per convention, we exclude effects with p >= .05. The three interfaces are compared to the baseline interface. The SEM model shows that all three interfaces are seen as more difficult than the baseline condition, with the exemplar + tag interface having roughly twice the impact (1.336) as the exemplar (0.655) and tag (0.649) interfaces. Both interfaces with exemplars are perceived more useful than the baseline interface, with the exemplar + tag interface almost twice as useful than the baseline as the exemplar interface. The tag interface is not more useful than the baseline (no significant loading). Reviewing the total structure of the model we observe that difficulty loads negatively on usefulness: the interfaces that

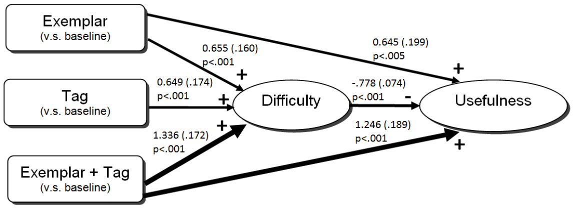


Figure 6.2: The structural equation model fitted on the Subject constructs difficulty and usefulness and their relations to the 4 interfaces

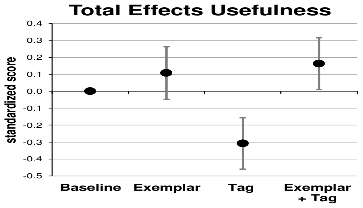


Figure 6.3: Total effects of the interfaces on usefulness: error bars are 1 standard error of the mean. Comparisons are made against the baseline interface.

[7] Hu and Bentler [99] propose cut-off values for the fit indices to be: CFI > .96, TLI > .95, and RMSEA < .05, with the upper bound of its 90% CI falling below 0.10

are more difficult are also perceived as less useful. We should therefore inspect the total effects on usefulness, which are plotted in figure 6.3. Compared to the baseline interface the two exemplar-powered interfaces are only slightly more useful (differences not significant) because their difficulty stands in the way of a much better user experience. This suggests that one could increase the user experience of these interfaces by making them less difficult (i.e less confusing, easier to understand), which should increase their perceived usefulness. Note that the total effect on usefulness suggests that the tag interface is actually perceived to be less useful than the baseline ($p < .05$) mainly because it is more difficult than the baseline. Thus the tag interface is seen as worse than the baseline interface.

6.4 Discussion

Our user survey reveals that in general providing rating support, as in the exemplar interface, helps users rate more consistently. Although our participants felt that the interfaces providing rating support are more difficult than our baseline interface, they liked these interfaces because they perceived the interfaces to be more useful. Several participants requested to see these interfaces implemented on MovieLens. This is consistent with our quantitative analyses. Between the two interfaces providing rating support, the exemplar one appears to have the lowest RMSE, the lowest minimum RMSE, and the least amount of natural noise.

Our quantitative analyses, however, show that reminding users of movie features in personalized fashion, as in the tag interface, did not help to improve the rating quality or the perceived usefulness of the interface. Our participants perceived the tag-interface as more difficult than the baseline. One potential reason for this is our tag selection algorithm. It is possible that an improved algorithm would select tags that have a more positive effect on the user.

The SEM analysis shows that the perceived usefulness and difficulty of our interfaces is related. This correlation is what led to having only small gains in total usefulness. While our interfaces had a positive effect on usefulness, they also had an opposite effect by being more perceived difficult. This leads us to believe that efforts simplifying our interfaces could significantly improve the user experience.

One of the contributions of this work is an exploration of a set of methodologies to evaluate rating interfaces. It is important that we have techniques for comparing novel recommender interfaces. Future work should continue exploring methodologies for comparing interfaces. Many of our metrics required a large amount of data to give statistically significant results, and metrics such as RMSE should be suspect when too few ratings are used. Likewise, while we feel the information theoretic metrics can be very informative, more work is still needed to develop robust techniques for comparing information theoretic values.

Improving prediction accuracy is one of the main interests in recommender system research, both in academia and in industry. For example, Netflix awarded $1 million dollars to a research team for the 10% of improvement in prediction accuracy[8]. Our efforts modifying recommender interfaces to gather high quality data with less noise complement the efforts to improve prediction accuracy. Overall we see the results of this work, and feedback from our participants, as promising. We believe that future work building upon our interfaces, algorithms, and methodologies can help us improve recommender systems.

[8] http://www.netflixprize.com

Chapter 7

Discussion and Acknowledgment

In this chapter, we summarize our contributions and discuss the implications of our work. We also reflect on the possible future directions based on our work.

7.1 Discussion

Through our work presented in this dissertation, we provide researchers and practitioners quantitative insights and evidence on how to improve the user experience. We focus on four aspects of the user experience:

- the long-term effect of a recommender system on users
- user satisfaction with the three fundamental properties of recommendations and recommendation lists
- the users' processes of mapping their preferences into rating scales
- the relations between different ways of measuring user experience.

These aspects cover three important areas: subsequent usage of recommendation taken, the properties of recommendations and recommendation lists, and the personalities of users. These areas have high impacts on user experience with recommender systems.

Users perceived recommender systems as helpful and useful when the systems suggest to them things they otherwise would not know. However, when investigating the long-term effect of a recommender system on users, we found that the recommender system exposed users to similar items over time. This finding begs the question of how much values users receive from recommender systems in the long run. We suggest that when a recommender system has a high tendency of providing similar items, users should be informed of how similar the recommendations are to their recent consumptions. Moreover, if users think the recommendations are very similar to what they usually consume, they should be able to adjust the recommendation lists to match their preferred level of similarity and diversity.

Although we show that recommender systems suggest similar items to users over time, we do not claim that this tendency of the systems is bad or good for an entire group of users. Researchers and practitioners have examined the levels of diversity, popularity, and serendipity that can enhance the experience of users as a group. Our results suggest that the perceived usefulness of recommender systems varies across users, and thus to improve user experience we need to personalize recommendations to individual users. Particularly, we find that users have different preferences for the diversity, popularity and serendipity of recommendations and recommendation lists. Besides the ability to adjust the diversity level of recommendations as discussed above, we should also allow users to adjust the level of popularity and serendipity. By increasing users' sense of control over what recommendations they receive, we can increase the perceived transparency of recommender systems about why certain items are suggested.

An even more exciting implication of our results is the way recommender systems can automatically adjust these three properties of recommendations (diversity, popularity, and serendipity) based on user personality. One of our findings is that users with high conscientiousness preferred serendipitous recommendations, but users with low conscientiousness did not. If the information about users' personalities is available, recommender systems may be able to automatically adjust the three properties accordingly.

In fact, although we propose to give users control over the three properties of recommendations, there are situations when automatic adjustment by the system is more beneficial to users, such as when users just finish the initial sign-in process. In this situation, users are new to the system, and thus may not fully understand all the controls, if any. The system, on the other hand, only has a few ratings for each user and is not able to infer their preferences for these three properties. Therefore, with information about the users' personalities, recommender systems can extrapolate individual preferences, and automatically adjust these recommendation properties for each user. The system can then take extra steps to explain what these properties mean to the users. We think this approach would improve the overall user experience, and solidify users trust in the system.

It is important to note that we do not advocate asking users personality questions before, or even after the sign up. Researchers have implemented several approaches to predict users' personalities via digital footprints. Recommender systems researchers and practitioners can apply these techniques to predict users' personalities, which, in turn, can be used to infer the preferences for diversity, popularity, and serendipity. Any system should provide users opt-out options if the users are not comfortable with any provided feature. Moreover, in a situation where a system does not have sufficient information of a new user, the system can take advantage of his/her digital footprints left on other rich platforms (such as Facebook, Twitter, or Pinterest, etc.). One way to acquire the digital footprints of a new user is to ask him/her to grant the system the assess to their information on the other social medias via API calls.

It's equally important to note that we do not advocate using user personality in every recommender system. For example, recommender system for political topics should always aim to deliver diversified information about different aspects of political policies so that users can have a better understanding of what is going on in their societies. Recommending to them information they already know, and limiting suggestions to viewpoints they already agree with, especially for controversial issues, could strengthen potential biases.

User experience is not only about the long-term effects of diminishing diversity nor about the three properties discussed above of recommendations and recommendation lists. User experience is also about how easily users can map their preferences onto a rating scale. Recommender systems in various domains employ different methods to collect signals about user preferences. Amazon[1] implicitly collects the signals by logging what users bought. Youtube[2] employs a binary rating scale to understand the preferences of their users. Netflix[3] and imdb[4] use a 5-star rating scale with a half-star increments. Facebook[5] provides users with a combination of a binary scale (like/dislike) and a set of emoticons to express their opinions. Users may have difficulties in mapping their preferences to the rating scale when switching from one scale to the other. Our research shows that we can facilitate the users' rating processes by providing some supports. These supports can be exemplars to remind users what they rated in the past, or tags to help them remember their impression about the items being rated. We provide quantitative evidence that providing these supports not only improved the user experience but also helped the system provide better recommendations.

Our research proposed several novel metrics and approaches to improve our understanding of the user experience with live recommender systems. For example, we introduce a new way to measure content diversity based on the information encoded in user-generated tags. Using this metric, instead of rating vectors, when examining user experiences with recommendation diversity can reduce bias in the results. Two movies with similar rating vectors that mean that they are similarly liked, not that their contents are similar. Researchers can use our novel metrics and approaches in their research about the user experience with recommender systems.

[1] https://www.amazon.com/

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

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

[4] http://www.imdb.com/

[5] https://www.facebook.com/

There are many challenges when doing research about user experience with recommender systems. First, understanding when we should evaluate user experience as a group or as individuals is important. We can consider the user experience for novel interfaces as the aggregated experience of a group of users because developing and maintaining a personalized interface for each user is costly and impractical. However, we can consider the user experience with a set of recommendations as individual experiences because the recommendation set is built upon individual's ratings, and developing and maintaining a personalized set is feasible. Second, we should consider other aspects of the user experience when incorporating user personality data into recommendation systems. For example, asking users personality questions to provide the necessary information for recommender systems to make personalized suggestions can improve user satisfaction. However, users can find this method intrusive and stop using a recommender system once it starts asking personality questions.

The limitation of this dissertation is that we have not looked at several other aspects of the user experience (such as trust and transparency). However, we hope our work inspires the next steps in the research to improve the user experience with recommender systems. In the next section, we list a few directions that can build upon our work.

7.2 Future work

Our work presents many opportunities for future research. For example, future work can integrate user personality into recommendation frameworks. This integration can lead to many promising research directions. One direction can be investigating what recommendation frameworks (e.g. SVD or item-item CF) can best improve the user experience when integrated with user personality. Another direction is to examine whether and how integrating personality can help users navigate through thousands, if not millions, of items. Vig et al. [14] presented a tag-genome information space that assists users in their navigation tasks. However, this tag-genome information space is not personalized. We wonder whether personalizing the tag-genome information space with user personality can improve the user experience when navigating through millions of items.

Even better, a recommender system with the information about user personality can determine how to minimize the filter bubble effect on recommendation consumption. For example, via user personality, the system can adjust the levels of diversity, popularity, and serendipity based on the preferences of the users. On the other hand, Harper et al. [100] reported that users would prefer to have mechanisms to tune the recommendations (e.g. the recommendation popularity) based on their preferences. Future work can investigate which of the two approaches, providing users some control or integrating user personality into recommendation frameworks, is better for the user experience.

Researchers can also examine the relationship between users' personalities and their recommendation system usage patterns. For example, introverted users tend to stay at home and watch more movies than extroverted users, and thus may be better at rating movies than extroverted users. Thus, a promising direction is to look at the rating quality of users with different personality types. If there are differences in the rating qualities, then the next step is to investigate how to incorporate these differences into recommender systems framework to improve the user experience.

Another future direction is to investigate how user personality influences the rating process. Prior work showed that asking users to re-rate items could lead to an improvement in recommendation quality. Thus, with user personality, we can ask whether highly conscientious users rate items more consistently, and thus do not need to re-rate items.

Our research results show that there is a broad spectrum of factors that affect the user experience with recommender systems. We hope our work will stimulate more research about user experience in recommender systems.

7.3 Acknowledgment

We are grateful for the support of the National Science Foundation, the University of Minnesota Informatics Institute, and colleagues in the Grouplens Lab during the completion of this dissertation. The work presented in chapter 3 was supported by the National Science Foundation grants IIS 08-08692 and IIS 10-17697. The late Professor John T. Riedl, our colleagues Shou Chang, Daniel Kluver and Aaron Halfaker contributed to the research in this chapter.

The work presented in chapter 5, and chapter 4 was supported by the National Science Foundation grant 1319382 and the University of Minnesota Informatics Institute 2015 UMII Informatics Fellowship.

The work presented in chapter 6 was supported by the National Science Foundation grants IIS 09-68483, IIS 08-08692, and IIS 10-17697.

References

[1] Samuel D Gosling, Peter J Rentfrow, and William B Swann. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528, 2003.

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

[3] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[4] Matthew R McLaughlin and Jonathan L Herlocker. A collaborative filtering algorithm and evaluation metric that accurately model the user experience. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 329–336. ACM, 2004.

[5] Joseph A Konstan and John Riedl. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2):101–123, 2012.

[6] Sean M McNee, John Riedl, and Joseph A Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06*

extended abstracts on Human factors in computing systems, pages 1097–1101. ACM, 2006.

[7] Conor Hayes and Padraig Cunningham. An on-line evaluation framework for recommender systems. *AH'2000 Workshop on Recommendation and Personalization in E-Commerce*, pages 50–59, 2002.

[8] Sean M McNee, Shyong K Lam, Joseph A Konstan, and John Riedl. Interfaces for eliciting new user preferences in recommender systems. In *User Modeling 2003*, pages 178–187. Springer, 2003.

[9] Sara Drenner, Shilad Sen, and Loren Terveen. Crafting the initial user experience to achieve community goals. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 187–194. ACM, 2008.

[10] Sean Michael Mcnee. *Meeting user information needs in recommender systems*. Proquest, 2006.

[11] Michael D Ekstrand, Praveen Kannan, James A Stemper, John T Butler, Joseph A Konstan, and John T Riedl. Automatically building research reading lists. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 159–166. ACM, 2010.

[12] Nava Tintarev and Judith Masthoff. Effective explanations of recommendations: user-centered design. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 153–156. ACM, 2007.

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

[14] Jesse Vig, Shilad Sen, and John Riedl. The tag genome: Encoding community knowledge to support novel interaction. *ACM Trans. Interact. Intell. Syst.*, 2(3):13:1 – 13:44, September 2012.

[15] Rong Hu and Pearl Pu. Enhancing collaborative filtering systems with personality information. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 197–204. ACM, 2011.

[16] Oliver P John and Sanjay Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138, 1999.

[17] Robert R McCrae and Paul T Costa. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81, 1987.

[18] Sam Gosling. *Snoop: What your stuff says about you.* Basic Books, 2009.

[19] Eli Pariser. *The Filter Bubble: What the Internet is Hiding from You.* Penguin, March 2012.

[20] Philip E Tetlock. *Expert political judgment: How good is it? How can we know?* Princeton University Press, 2005.

[21] Cass R. Sunstein. *Republic.com: XA-GB. ...* Princeton University Press, 2002.

[22] Nicholas Negroponte. *Being Digital.* Random House LLC, January 1996.

[23] Nicholas Negroponte. 000 000 111 - double agents. http://www.wired.com/wired/archive/3.03/negroponte_pr
.html, visited on 2013-09-13.

[24] Greg Linden. Eli pariser is wrong. http://glinden.blogspot.com/2011/05/eli-pariser-is-wrong.html, visited on 2013-09-13.

[25] Tomonari Kamba, Krishna A Bharat, and Michael C Albers. The krakatoa chronicle-an interactive, personalized newspaper on the web. 1995.

[26] Daniel Fleder and Kartik Hosanagar. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management Science*, 55(5):697–712, May 2009.

[27] Kartik Hosanagar, Daniel M. Fleder, Dokyun Lee, and Andreas Buja. Will the global village fracture into tribes: Recommender systems and their effects on consumers. SSRN Scholarly Paper ID 1321962, Social Science Research Network, Rochester, NY, October 2012.

[28] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32. ACM, 2005.

[29] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 13–22. ACM, 2012.

[30] Òscar Celma and Pedro Cano. From hits to niches?: or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, page 5. ACM, 2008.

[31] Kamal Ali and Wijnand Van Stam. Tivo: making show recommendations using a distributed collaborative filtering architecture. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 394–401. ACM, 2004.

[32] Ronald W Shephard and Rolf Färe. *The law of diminishing returns.* Springer, 1974.

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

[34] Daniel M Fleder and Kartik Hosanagar. Recommender systems and their impact on sales diversity. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 192–199. ACM, 2007.

[35] Gediminas Adomavicius and YoungOk Kwon. Toward more diverse recommendations: Item re-ranking methods for recommender systems. In *Workshop on Information Technologies and Systems*, 2009.

[36] Harald Steck. Item popularity and recommendation accuracy. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 125–132. ACM, 2011.

[37] Gerbert Kraaykamp and Koen Van Eijck. Personality, media preferences, and cultural participation. *Personality and individual differences*, 38(7):1675–1688, 2005.

[38] Peter J Rentfrow and Samuel D Gosling. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology*, 84(6):1236, 2003.

[39] Olivia Chausson. Who watches what?: assessing the impact of gender and personality on film preferences, 2010.

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

[41] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pages 677–686. ACM, 2014.

[42] Jinoh Oh, Sun Park, Hwanjo Yu, Min Song, and Seung-Taek Park. Novel recommendation based on personal popularity tendency. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 507–516. IEEE, 2011.

[43] Mi Zhang and Neil Hurley. Novel item recommendation by user profile partitioning. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 508–515. IEEE Computer Society, 2009.

[44] Saúl Vargas and Pablo Castells. Exploiting the diversity of user preferences for recommendation. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 129–136. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2013.

[45] Michael D Ekstrand, F Maxwell Harper, Martijn C Willemsen, and Joseph A Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 161–168. ACM, 2014.

[46] Nava Tintarev and Judith Masthoff. A survey of explanations in recommender systems. In *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*, pages 801–810. IEEE, 2007.

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

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

[49] Judith S Olson and Wendy A Kellogg. *Ways of Knowing in HCI*. Springer, 2014.

[50] Jilin Chen, Werner Geyer, Casey Dugan, Michael Muller, and Ido Guy. Make new friends, but keep the old: recommending people on social networking sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 201–210. ACM, 2009.

[51] Joeran Beel, Stefan Langer, Marcel Genzmehr, Bela Gipp, Corinna Breitinger, and Andreas Nürnberger. Research paper recommender system evaluation: a quantitative literature survey. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, pages 15–22. ACM, 2013.

[52] Will Hill, Larry Stead, Mark Rosenstein, and George Furnas. Recommending and evaluating choices in a virtual community of use. In *In Proc. of CHI '95*, pages 194–201, New York, NY, USA, 1995.

[53] G. Adomavicius, J. Bockstedt, S. Curley, and J. Zhang. Recommender systems, consumer preferences, and anchoring effects. In *Decisions@RecSys Workshop*, pages 35–42, Chicago, 2011.

[54] Dan Cosley, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, and John Riedl. Is seeing believing?: how recommender system interfaces affect users' opinions. In *In Proc. of CHI '03*, pages 585–592, New York, NY, USA, 2003. ACM.

[55] Michael P. O'Mahony, Neil J. Hurley, and Gunol C.M. Silvestre. Detecting noise in recommender system databases. In *Proceedings of the 11th international conference on Intelligent user interfaces*, IUI '06, pages 109 – 115, New York, NY, USA, 2006. ACM.

[56] Xavier Amatriain, Josep M. Pujol, Nava Tintarev, and Nuria Oliver. Rate it again: increasing recommendation accuracy by user re-rating. In *In Proc. of RecSys '09*, pages 173 – 180, New York, NY, USA, 2009. ACM.

[57] Daniel Kluver, Tien T. Nguyen, Michael Ekstrand, Shilad Sen, and John Riedl. How many bits per rating? In *In Proc. of RecSys '12*, pages 99 – 106, New York, NY, USA, 2012. ACM.

[58] E. Isaac Sparling and Shilad Sen. Rating: how difficult is it? In *In Proc. of RecSys '11*, pages 149 – 156, New York, NY, USA, 2011. ACM.

[59] Sean M. McNee, Shyong K. Lam, Joseph A. Konstan, and John Riedl. Interfaces for eliciting new user preferences in recommender systems. In Peter Brusilovsky, Albert Corbett, and Fiorella de Rosis, editors, *User Modeling 2003*, number 2702 in Lecture Notes in Computer Science, pages 178–187. Springer Berlin Heidelberg, January 2003.

[60] Itamar Simonson. Determinants of customers' responses to customized offers: Conceptual framework and research propositions. *Journal of Marketing*, 69(1):32–45, January 2005.

[61] James R. Bettman, Mary Frances Luce, and John W. Payne. Constructive consumer choice processes. *Journal of Consumer Research*, 25(3):187 – 217, December 1998. ArticleType: research-article / Full publication date: December 1998 / Copyright ©1998 Journal of Consumer Research Inc.

[62] Baruch Fischhoff. Value elicitation: Is there anything in there? *American Psychologist*, 46(8):835–847, 1991.

[63] Sarah Lichtenstein and Paul Slovic. *The Construction of Preference.* Cambridge University Press, September 2006.

[64] E. U. Weber and E. J. Johnson. Mindful judgment and decision making. In *Annual Review of Psychology*, volume 60, pages 53–85. Annual Reviews, Palo Alto, 2009.

[65] Matt Marshall. Aggregate knowledge raises $5m from kleiner, on a roll | VentureBeat. http://venturebeat.com/2006/12/10/aggregate-knowledge-raises-5m-from-kleiner-on-a-roll/, visited on 2013-09-06.

[66] Xavier Amatriain and Justin Basilico. The netflix tech blog: Netflix recommendations: Beyond the 5 stars (part 1). http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html, visited on 2013-09-06.

[67] Sylvain Senecal and Jacques Nantel. The influence of online product recommendations on consumers" online choices. *Journal of Retailing*, 80(2):159–169, 2004.

[68] Bo Xiao and Izak Benbasat. E-commerce product recommendation agents: use, characteristics, and impact. *MIS Q.*, 31(1):137?209, March 2007.

[69] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. GroupLens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, CSCW '94, pages 175–186, New York, NY, USA, 1994. ACM.

[70] Marshall Van Alstyne and Erik Brynjolfsson. Global village or cyber-balkans? modeling and measuring the integration of electronic communities. *Management Science*, 51(6):851–868, 2005.

[71] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA, 2001. ACM.

[72] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[73] F Maxwell Harper, Xin Li, Yan Chen, and Joseph A Konstan. An economic model of user rating in an online recommender system. In *User Modeling 2005*, pages 307–316. Springer, 2005.

[74] Jesse Vig, Shilad Sen, and John Riedl. Navigating the tag genome. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 93–102. ACM, 2011.

[75] Tien T. Nguyen, Daniel Kluver, Ting-Yu Wang, Pik-Mai Hui, Michael D. Ekstrand, Martijn C. Willemsen, and John Riedl. Rating support interfaces to improve user experience and recommender accuracy. *To appear in the seventh ACM Recommender System Conference, RecSys 2013*, October 2013.

[76] Paul T Costa and Robert R McCrae. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2:179–198, 2008.

[77] Anthony E Kemp. *The musical temperament: Psychology and personality of musicians.* Oxford University Press, 1996.

[78] Shuo Chang, F. Maxwell Harper, and Loren Terveen. Using groups of items for preference elicitation in recommender systems. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*, CSCW '15, pages 1258–1269, New York, NY, USA, 2015. ACM.

[79] Jesse Vig, Shilad Sen, and John Riedl. The tag genome: Encoding community knowledge to support novel interaction. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):13, 2012.

[80] Marko Tkalcic, Matevz Kunaver, Jurij Tasic, and Andrej Košir. Personality based user similarity measure for a collaborative recommender system. In *Proceedings of the 5th Workshop on Emotion in Human-Computer Interaction-Real world challenges*, pages 30–37, 2009.

[81] Rong Hu and Pearl Pu. Acceptance issues of personality-based recommender systems. In *Proceedings of the third ACM conference on Recommender systems*, pages 221–224. ACM, 2009.

[82] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 180–185. IEEE, 2011.

[83] Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. Personality and patterns of facebook usage. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 24–32. ACM, 2012.

[84] Wu Youyou, Michal Kosinski, and David Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040, 2015.

[85] Alexandros Karatzoglou, Xavier Amatriain, Linas Baltrunas, and Nuria Oliver. Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 79–86. ACM, 2010.

[86] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 452–461. AUAI Press, 2009.

[87] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53, 2010.

[88] Alan Said, Brijnesh J. Jain, Sascha Narr, and Till Plumbaum. Users and noise: The magic barrier of recommender systems. In *UMAP 2012*, pages 237–248. Springer, 2012.

[89] Dirk Bollen, Mark Graus, and Martijn C. Willemsen. Remembering the stars?: effect of time on preference retrieval from memory. In *In Proc. of RecSys '12*, pages 217 – 220, New York, NY, USA, 2012. ACM.

[90] Tien T. Nguyen and John Riedl. Predicting users' preference from tag relevance. *User Modeling, Adaptation, and Personalization, UMAP 2013*, pages 274–280, 2013.

[91] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, September 1974.

[92] Christopher K. Hsee. The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, 67(3):247–257, September 1996.

[93] Amos Tversky, Shmuel Sattath, and Paul Slovic. Contingent weighting in judgment and choice. *Psychological Review*, 95(3):371–384, 1988.

[94] Syavash Nobarany, Louise Oram, Vasanth Kumar Rajendran, Chi-Hsiang Chen, Joanna McGrenere, and Tamara Munzner. The design space of opinion measurement interfaces: exploring recall support for rating and ranking. In *In Proc. of CHI '12*, pages 2035 – 2044, New York, NY, USA, 2012. ACM.

[95] Thomas Mussweiler. Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review*, 110(3):472–489, 2003.

[96] F. Maxwell Harper, Xin Li, Yan Chen, and Joseph A. Konstan. An economic model of user rating in an online recommender system. In *User Modeling 2005*, pages 307–316. Springer, 2005.

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

[98] Michael D. Ekstrand, Michael Ludwig, Joseph A. Konstan, and John T. Riedl. Rethinking the recommender research ecosystem: reproducibility, openness, and LensKit. In *In Proc. of RecSys '11*, pages 133 – 140, New York, NY, USA, 2011. ACM.

[99] Li-tze Hu and Peter Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55, 1999.

[100] F. Maxwell Harper, Funing Xu, Harmanpreet Kaur, Kyle Condiff, Shuo Chang, and Loren Terveen. Putting users in control of their recommendations. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, pages 3–10, New York, NY, USA, 2015. ACM.