Leveraging Social Information To Improve Recommendation Novelty

Ishan Ghanmode
4621735



Leveraging Social Information To Improve Recommendation Novelty

Master's Thesis in Computer Science

Web Information Systems Group
Faculty of Electrical Engineering, Mathematics, and Computer Science
Delft University of Technology

Ishan Ghanmode
4621735

23rd August 2018



Author
Ishan Ghanmode (4621735)

Title
Leveraging Social Information To Improve Recommendation Novelty

MSc presentation
31st August 2018

Graduation Committee

Prof. dr. ir. G. J. P. M. Houben	TU Delft, Full Professor (Chair)
Dr. Nava Tintarev	TU Delft, Assistant Professor & Supervisor
Dr. Cynthia Liem	TU Delft, Assistant Professor

Abstract

In today's digital world, users are often confronted with an abundance of information. Whether the user is looking to compare online prices for products, searching for new movies to watch or music to listen, the available information at hand exceeds the amount of information which the user wants to consider before making a choice. For this, various recommendation systems have been developed. Primarily, the recommendations made by these recommendations systems have been evaluated based on their accuracy. More recent research has begun evaluating the subjective perception of the recommendations and has developed additional attributes such as *diversity* and *novelty* to evaluate recommendation systems. However, the impact of such attributes on actual *user satisfaction* has been explored less. Recent research has seen an increase in evaluating other aspects of recommendation systems such as recommendation interfaces. Nonetheless, very limited work has been done in terms of recommendation system interfaces to improve the perceived quality of the recommendations and overall *user satisfaction*.

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

Preface

"Dream, dream, dream. Dreams transform into thoughts and thoughts result in action." These golden words by Dr. A. P. J. Abdul Kalam have always instilled vigor and a sense of excitement in the endeavors I have undertaken throughout my life. This master thesis, probably one the most challenging, yet an enlightening and educational journey gave me an excellent opportunity to work on some of my most favorite domains in Computer Science.

Many people have supported me in this endeavor whom I would like to express my gratitude. First and foremost, I would like to thank my supervisor Dr. Nava Tintarev for giving me this fantastic opportunity to work with her on this exciting topic. This work would not have been possible without her invaluable guidance and insight. Her unrivaled enthusiasm, work ethic and "thirst for knowledge" attitude is inspiring. It kept me motivated and pushed me to work harder. I would like to thank all the thesis committee members for assessing and evaluating my work.

Most importantly, my parents. I thank them for supporting me through the thick and thin, and for teaching me the value of education and working hard. Their love and support is what kept me going all this time. I hope with this work, I made them proud.

My sincere gratitude to the members of Epsilon - for their valuable feedback on the thesis and presentations. Their continuous feedback in the past few months helped me shape my thesis. Furthermore, I would like to express my gratitude to all my friends across the globe for their valuable inputs.

Finally, my eternal gratitude and love to Sameera. Thank you for being my pillar of strength during the tough times and giving me all the love, inspiration and encouragement.

Ishan Ghanmode

Delft, The Netherlands
23rd August 2018

Contents

Preface v

1 Introduction 1
- 1.1 Problem Statement . . . 2
- 1.2 Research Objectives . . . 3
- 1.3 Contributions . . . 3
- 1.4 Thesis Outline . . . 4

2 Background and Related Work 5
- 2.1 Introduction to Social Recommendation Systems . . . 5
- 2.2 Definition of Social Recommendation . . . 6
- 2.3 Existing Social Recommendation Systems . . . 7
  - 2.3.1 Overview of the Collaborative Filtering Process . . . 8
  - 2.3.2 Memory Based Systems . . . 9
  - 2.3.3 Model Based Systems . . . 9
  - 2.3.4 Discussion . . . 10
- 2.4 Inspectability and Control in Recommendation Systems . . . 12
  - 2.4.1 Inspectability . . . 12
  - 2.4.2 Control . . . 14
  - 2.4.3 Impact of Recommendation System Interfaces on Users . . . 15
- 2.5 Evaluation . . . 16
  - 2.5.1 Beyond Accuracy Metrics . . . 17
  - 2.5.2 Evaluation of Social Recommendation systems . . . 20
- 2.6 Conclusions . . . 20
- 2.7 Gaps and Thesis Motivation . . . 21
  - 2.7.1 Gaps . . . 21
  - 2.7.2 Thesis Motivation . . . 21

3 System 23
- 3.1 System Overview . . . 23
  - 3.1.1 System Design . . . 25
- 3.2 Variable Description . . . 26
  - 3.2.1 Independent Variables . . . 26

3.2.2 Dependent Variables . . . 26
3.3 Initial Recommendations Phase . . . 27
3.3.1 MovieLens 1M Dataset . . . 27
3.3.2 On-boarding of New Users . . . 29
3.3.3 Initial Seed Items . . . 30
3.3.4 Recommendation Algorithm . . . 35
3.3.5 Collecting User Demographics . . . 36
3.3.6 Process Flow . . . 36
3.4 Social Information Phase . . . 41
3.4.1 MovieTweetings Dataset . . . 41
3.4.2 Process Flow . . . 46
3.5 Revised Recommendations Phase . . . 53
3.5.1 Selection of Relevant Movies . . . 53
3.5.2 Maximal Marginal Relevance . . . 54
3.5.3 Process Flow . . . 55

4 Evaluation **57**
4.1 Approach . . . 57
4.1.1 Experimental Setup . . . 58
4.1.2 Procedure . . . 58
4.2 Evaluation Metrics . . . 59
4.2.1 Offline Measures . . . 59
4.2.2 Online Measures . . . 60
4.3 Research Objectives . . . 62
4.4 Results . . . 63
4.4.1 Participants . . . 63
4.4.2 Data Collection . . . 63
4.4.3 Hypotheses Results . . . 71
4.4.4 Discussion . . . 75

5 Discussion **79**
5.1 Impact of System Diversity on Discovery Perception . . . 80
5.2 Impact of Discovery Perception on User Satisfaction . . . 81
5.3 Limitations . . . 81

6 Conclusion **83**
6.1 Future Work . . . 84

A User Demographics Questionnaire **93**

B Result Data **95**
B.1 Perceived Novelty . . . 95
B.2 Perceived Diversity . . . 97
B.3 User Satisfaction . . . 98

B.4 System Novelty 99
B.5 System Diversity 100
B.6 Popularity Scores 101

Chapter 1

Introduction

Recommendation systems play an increasingly important role in different online applications. These online applications are characterized by very large amounts of data such as news, music, images or movies. Their goal is to filter relevant information and present items that are most likely of interest to them. Traditionally, the field of recommendation systems has evaluated the recommendations using traditional metrics of accuracy and precision.

In recent years, researchers have come to realize that the goal of a recommendation system is not only restricted to making accurate predictions. They extend to providing personalized help in helping users discover new and relevant items. This caused two major changes in the field.

First, researchers in [1] stated that being just "accurate" was not good enough for recommendations. They needed to be studied from a "user-centric perspective" to make them not only accurate and helpful, but also fun to use. The researchers also suggested to broaden horizons while evaluating recommendations. This suggestion gave rise to research that evaluated recommendation systems in an online scenario with user-centric evaluation metric such behaviors (consumption behavior) as well as attitudes (choice satisfaction, usefulness and so on). User-centric evaluation approaches to recommendation systems have received interest in the research and industry domains of recommendation system and HCI communities. Secondly, the researchers stated that aside from the algorithmic aspect of recommendation systems, the interactive components of a recommendation system should also be investigated. The quality of the recommendations along with the impact of the interfaces should collectively be studied to understand the system impact on the users. These two changes prompted a broader research perspective in the field of recommendation systems.

This study is based on the above two research directions. In the next sections we discuss more about the problem in detail, our research objectives, our contributions and the structure of this report.

1.1 Problem Statement

Users today are often confronted with an abundance of information. Whether the user is looking to compare online prices for products, searching for new movies to watch or music to see, the available information at hand exceeds the amount of information which the user wants to consider before making a choice. For this, various recommendation systems have been developed. Some notable examples are Netflix[1] for movie recommendations, Amazon[2] for recommending items to purchase. Characterizing and evaluating the quality of user experience and user's subjective attitudes towards the quality of recommendations is an important issue which draws attention from researchers. This is because the recommender technology is becoming widely accepted as an important component which provides both user benefits and improvement in the business revenue. This is summarized as:

- Evaluating user's perception of the recommendations can help researchers and developers understand the level of satisfaction and the intended benefits.
- Understanding the satisfaction levels of the users helps improve the various aspects of the system such as recommendation algorithm, interface capabilities and so on.

Recent years has seen the rise of a new kind of recommendation system called social recommendation systems. Social networks such as Facebook and Twitter have emerged as some of the most popular social media platforms allowing users to communicate and express their opinions and feedback. This online interaction between various users generates a lot of social information which could be effectively harnessed in areas such as recommendation systems. Social recommendation systems harness this social information to make recommendations.

To summarize the current recommendation systems research:

- Primarily, previous research has been focused on evaluating the recommendations generated based on their accuracy. More recent research has began evaluating the subjective perception of the recommendations and have developed additional attributes such as *diversity* and *novelty* to evaluate recommendation systems.
- Much focus has been on making better algorithms for improving the quality of recommendations. Recent research has seen increase in evaluating other aspects of recommendation systems such as recommendation interfaces.

We base this study on the following problems at hand:

- Limited research into how interactive interfaces in social recommendation systems impact the system and perceived quality of recommendations.

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

- Not enough research has been done in the field to understand and correlate the impact of perceived attributes of recommendations such as *diversity* and *novelty* on overall *user satisfaction*.

1.2 Research Objectives

Based on the problem statement defined above, we have two main research questions which form the basis of this study.

RQ1: Does incorporating social information with an interactive interface help users discover novel and diverse items?

Objective: To answer this question, we aim to build an interactive social recommendation system interface. We plan to generate recommendations based on the choices made by the user and analyze the quality of the recommended items. We plan to measure different system-level and perceived quality metrics of the recommendations.

RQ2: Does discovering novel and diverse items help achieve better levels of user satisfaction?

Objective: To answer this question, we aim to conduct an online user experiment to gather the relevant user data. Then, we aim to study & measure the impact of the recommended items on the satisfaction levels of the users.

1.3 Contributions

Our main contributions have been the following:

- We analyze the current gaps in the interface and evaluation aspects of the recommendation system domain.
- We introduce an interactive movie recommendation system called *MovieTweeters*. It combines aspects of interactive interfaces with recommendation algorithms to generate recommendations for users. We analyze the impact of the perceived quality of the recommendations on the overall satisfaction of the users.
- We conducted an user experiment of our *MovieTweeters* interface. Our initial results show the following:
  - Users discovered more *novel* content interacting with their relevant social information.

– The perceived quality of the recommended movies had a direct impact on the user satisfaction. As users perceived their movies to be more *novel* and *diverse*, their satisfaction levels also increased.

1.4 Thesis Outline

We begin with an in-depth analysis of the background work done in the field in Chapter 2. In Chapter 3 we give a formal description of our system model and architecture. Next, we describe our evaluation metrics and experimental setup in Chapter 4. We discuss our obtained results in Chapter 5 and finally we conclude in Chapter 6 and discuss future work.

Chapter 2

Background and Related Work

To answer our *research questions* dealing with the impact of social information on the quality of recommendations and the impact of perceived quality attributes of recommendations on user satisfaction (as stated in Chapter 1), we first look in depth the background and related work in the Social Recommendation Systems domain. We aim to answer our research questions with the help of an interactive interface which would allow the users inspect the system controls and help them discover new items. For this, we also study the past work done in the field of inspectability and control, beyond accuracy metrics and evaluation of social recommendation systems.

2.1 Introduction to Social Recommendation Systems

Recommendation systems, with the advent of the World Wide Web, play a significant role in helping online users find new items potentially of interest to them. Consumers often want to be recommended the best possible products and items for their requirements. Various types of such recommendation systems have been implemented in different domains such as e-commerce (Amazon[1]) and entertainment (Netflix[2]). There are different techniques which are used to build such recommendation systems. They can broadly be classified into content-based methods, collaborative filtering based methods [2] and hybrid methods [3] [4]. Content-based methods recommend relevant items to the users based on their previous consumption of items. Collaborative filtering based methods recommend items to a user from other users with similar interests and behavior [5]. Hybrid based methods not only combine both the content-based and CF-based methods.

Recent years has seen the surge in social media such as social networking websites. This increasing popularity of various social media platforms has encouraged user's activities with their friends, families, and with other social media users. This

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

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

rising activity produces different social relations such as friendships on Facebook[3], followers on Twitter[4] and so on. Social platforms like these also provide the users with an opportunity to express their views and opinions about a number of subjects. According to social correlations theories such as social influence [6], a user's preference can be influenced by their socially close friends or other close users of similar interests. These social media platforms and social relations could be potentially exploited to improve the performance of recommendation systems.

2.2 Definition of Social Recommendation

Social networks such as Facebook and Twitter have emerged as some of the most popular social media platforms allowing users to communicate and express their opinions and feedback. This online interaction between various users generates a lot of social information which could be effectively harnessed in areas such as recommendation systems. For instance, Facebook has more than 1 billion users making it one of the first social networking to do so and Twitter averaging around 360 million active users per month. This rapid growth of social media has encouraged the development of social recommendation systems [7]. This section discusses the basis and definition of social recommendations and how are they different from traditional recommendations.

One of the earliest works which included social properties was done by Kautz et al. in their work [8] who built "ReferralWeb" which was an interactive system for searching relevant social networks on the World Wide Web. One of the definitions for social recommendation as defined by the authors in [9] is any given recommendation with online social information as an additional input i.e augmenting or improving the existing social recommendations with additional social information. Social information can be in the form of social relations, friendships, social influence and so on. In this definition, the social recommendation systems assume that the users are related when they establish social relations. Under this assumption, the social information or social relations are used to improve the performance of the recommendations. Some social recommendation systems which form under this definition are the following systems: [10], [11], [12], [13], [14].

Another definition of social recommendation can be based on a much broader perspective [15]. Here, the social recommendations are recommendations which include items in the social media domain, user communities [16] and tags [17]. The data sources which the social recommendation systems use are not only limited to online social relations (as stated above), are also use different social media items such as social tagging [18], user interactions [19], user click behaviors and so on. Kofler, Christoph, et al. in [20] state that a social media based platform incorporating user-contributed meta-data can be used for providing more authentic and personalized recommendations.

[3] https://www.facebook.com
[4] https://www.twitter.com

2.3 Existing Social Recommendation Systems

In this section, the existing social recommendation systems across different domains are discussed. Most of the existing social recommendation systems adopt collaborative filtering (CF) techniques [21] [22]. Here, social recommendation consists of two main inputs: 1. user-item rating information, 2. social information. The basic social recommendation model can be defined as the combination of a basic collaborative filtering model and a social model. For our study, we use this model to build our system for answering our *research questions*. Authors in [9], divide the existing social recommendation systems into two types of CF models: memory based systems and model based systems. Memory based systems uses the data available to establish correlation between users and items to recommend unknown items to the users. Memory based systems rely heavily on given similarity measures to match similar users or similar items. Memory based techniques are not always fast and scalable. To achieve more faster and scalable real time recommendations based on large datasets, model based systems were developed. These systems use a number of machine learning techniques to train a vector of items for a specific user, which can result in a model that can predict user's rating for a new item. We now discuss a brief overview of the Collaborative Filtering Process.

Table 2.1 classifies the different existing social recommendation systems.

Reference	Type of System	Model	Domain	Dataset
[13]	Memory Based	Random Walk Model	Product Reviews	Epinions
[23]	Memory Based	Trust Aware Recommendation System	Product Reviews	Epinions
[24]	Memory Based	Trust Based Model	Product Reviews	Epinions
[10]	Memory Based	Trust Based ACF Model	Movies	FilmTrust
[25]	Memory Based	Collaborative Filtering Model	Movies	NA
[26]	Model Based	Probabilistic Matrix Factorization Model	Product Reviews	Epinions
[27]	Model Based	Co-factorization Machine Model	Micro blogging	Twitter
[28]	Model Based	Matrix Factorization Model	Online Communities	Flixster, Epinions
[14]	Model Based	Matrix Factorization Model	Online Communities	Douban, Epinions
[16]	Model Based	ARM & LDA Models	Micro blogging	Orkut
[29]	Model Based	Graph Based Model	Music	Last.fm
[30]	Model Based	Graph Based Model	Product Reviews	Epinions
[12]	Model Based	SVM Model	Product Reviews	Epinions

Table 2.1: Classification of Social Recommendation Systems.

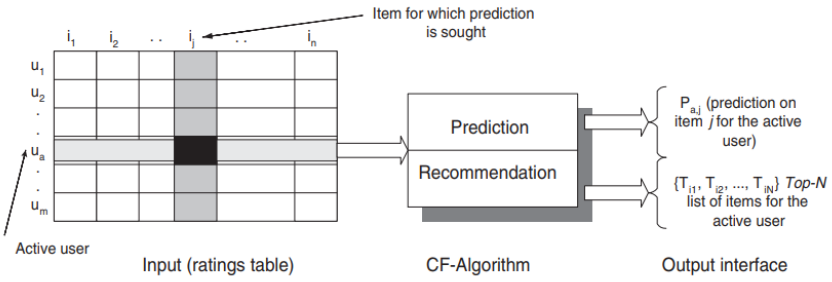


Figure 2.1: Collaborative Filtering [22]

2.3.1 Overview of the Collaborative Filtering Process

The goal of a CF algorithm is to recommend new items to users based on the user's previous item preferences and opinions of other like-minded users. Consider a typical CF scenario, there is a list of *m* users, $U = \{u_1, u_2, u_3, u_4.....u_m\}$ and a list of *n* items, $I = \{i_1, i_2, i_3, i_4.....i_n\}$. Each user u_i expresses his/her opinions about a list of items, I_{u_i}. There are two ways or expressing such opinions,

- *Explicitly:* Opinions could be given by the users as a rating score, usually within a certain numerical scale.

- *Implicitly:* Opinions could be derived implicitly from analyzing user records, time logs, click behavior and so on.

There exists an *active user* $u_a \in U$. The task here for a CF algorithm is to find an item likeliness which can be derived in two forms:

- *Making predictions:* A prediction value is a numerical value $P_{a,j}$, which predicts the likeliness of an item $i_j \notin I_{u_a}$ for the active user u_a. This value is within the same scale as the opinion values provided by the u_a.

- *Top-N Recommendation list:* This is a list of *N* items which the active user will like the most. Note, that the recommended list must have items that the active user has not already consumed.

Figure 2.1 shows how the working of a CF algorithm. CF algorithms represent user-item data as a ratings matrix, *A*. Each entry $a_{i,j}$ in *A* represents the ratings of the *i*th user on the *j*th item. All the ratings within a numerical scale. 0 represents no interaction between the user and the particular item.

Next, categorization of the existing social recommendation systems is done along with the domains and the datasets which have been used respectively.

2.3.2 Memory Based Systems

As stated above, many of the social recommendation systems use collaborative filtering models. In memory based models, missing rating for given users are aggregated from the ratings of users correlated to them. An additional input is used in the form of social information. Different systems use different techniques to achieve this.

Quite some work has also been done on the concept of *trust* along with the social information aspect to build better recommendation systems. *Trust* has different definitions in the context of recommendation systems research. It can be defined as the user's confidence in the recommendation system, and so on. Authors of [13] developed a system called *TrustWalker* which stressed on the importance of trust-based approaches along with the capabilities of item-oriented approaches. *TrustWalker* suggested a random walk model which combined the trust and user based approaches into one framework. It queried a user's direct and indirect friend's ratings for a target item as well as similar items by the process of random walks on online social networks. Massa and Avesani in [23] [24] propose a trust aware recommendation system to overcome problems that collaborative filtering systems face such as cold start problem (with new users), data sparseness and unreliable users. The authors also discuss the implementation of web of trust, allowing users to rate other users in terms of trust. Their system takes the trust matrix along with the ratings matrix and produces an output matrix of predicted ratings that the users would assign to the items. They tested their framework on epinions.com, an online community.

In the movie domain, Golbeck in her work [10] designed an interface *FilmTrust*, a website which uses trust in social networks to predict movie recommendations for users. *TidalTrust*, a trust predictive algorithm, was deployed in the system and used as a basis to generate predictive ratings for users. The users were first asked to add a list of friends in their network and assign them trust values. The second half of the interface deals with rating certain movies. The users were asked to rate the movies and were also show predictive ratings from their most trusted friends. Basu et. al in their work [25] developed an approach to exploit artifact information and social content information to make movie recommendations. Traditionally, only user ratings along with social information is used to make predictions. The authors propose a learning approach which not only looks at the ratings, but also other forms of information about the artifact (item) to predict user preferences.

2.3.3 Model Based Systems

Model based social recommendation systems choose model based collaborative filtering methods as their underlying models. Matrix factorization has been widely used in such systems. Many existing social recommendation systems use matrix factorization techniques. Ma et al. in their work [26] propose a system, *SoRec* based on a factor analysis approach performed on probabilistic matrix factorization

to solve problems of data sparsity and bad prediction accuracy by utilizing both the user's social network information and rating information. They connect the two data sources through a shared user latent space. By performing factor analysis, the required low-rank user latent features and item latent feature space are calculated to make the relevant social recommendations. This can also be described as a co-factorization method. The underlying assumption is that the *i-th* user u_i should have similar user preference vector u_i in the rating and social information space. In another work done by Hong et. al in [27], tackle the issue of predicting user decisions and modeling user's interests in Twitter by proposing co-factorization machines (CoFM).

Jamali and Ester in their work [28], incorporate trust propagation in the matrix factorization model to make social network based recommendations. Their model, *SocialMF* achieved reduced recommendations error particularly for cold start users. They based their approach on regularization methods where the main focus is on a user's preference. It assumes that a user's preference is similar to that of their social network. On the contrary, Ma et al. in their work [14] differentiate between trust aware recommendations systems and social recommendations. They state that trust aware recommendation systems are based more on the assumption that users have similar taste with the users they trust is not always true. This is because in social recommendation systems taste of one user's friends may vary very significantly. The authors proposed social regularization, keeping in mind that users in social network may have diverse tastes.

In other model based techniques, in [16] Chen et al. work on two algorithms namely, association rule mining (ARM) and Latent Dirichlet Allocation (LDA) to discover and recommend new communities in order to meet new users on Orkut. Their results show that, LDA performs better than ARM when recommending four or more than four communities. In graph based approaches, explicit social relationships have been studied to provide better social recommendations. Zhao et al. in [29] study the explicit social relationships on social network and develop an framework using a novel graph approach to fuse social relationships with the item rating matrix and produce neighborhood similarities for recommendations. Building on similar graph based approaches, Symeonidis et al. in [30] work on Social Rating Networks (SRNs) such as Epinions.

2.3.4 Discussion

Social relations can be used as reliable and independent sources for making recommendations. As seen above, different approaches are used to build social recommenders. Researchers have used both user and item based collaborative filtering approaches to build such recommenders. An overview and segregation of the existing social recommendation systems according to the approaches has been done to provide the reader with an quick overview of the systems in various different domains. Increasing popularity of social medial platforms result in different online relations between users. Traditional recommendation systems form the hypothesis

that users are independent and are identically distributed. On the other hand, social users are connected to each other in different relations such as friendships (Facebook) and followers (Twitter) as shown in Figure 2.2.

	v_1	v_2	v_3	v_4	v_5
u_1	5	?	2	?	?
u_2	4	4	5	?	?
u_3	?	?	4	4	1
u_4	?	?	?	?	3
u_5	?	?	1	?	?

(b) Traditional Recommender Systems

	v_1	v_2	v_3	v_4	v_5
u_1	5	?	2	?	?
u_2	4	4	5	?	?
u_3	?	?	4	4	1
u_4	?	?	?	?	3
u_5	?	?	1	?	?

	u_1	u_2	u_3	u_4	u_5
u_1	0	1	0	0	1
u_2	0	0	1	1	0
u_3	0	0	0	0	0
u_4	1	0	1	0	0
u_5	0	1	1	1	0

(c) Social Recommender Systems

Figure 2.2: Traditional vs Social Recommendation Systems [9]

In the real world, people often seek their friends and family for recommendations. This has an analogy with many online communities where users ask other users for new and better recommendations. Some of the advantages of social recommendations from the literature are summed up below:

- A user's preferences are more likely to be similar to users in their own social network rather that those of randomly chosen users. Connected users can provide good recommendations.
- Social recommendations can improve the coverage and diversity of recommendations.
- Different user's opinions and tastes can be understood via the social data such as social media tags, social relations on social networking websites and so on.

In summary, some of the negatives of social recommendations can be summarized as the following:

- Social data can be noisy and may have a negative impact on the recommendations.
- There can be a scenario where cold-start users may have little social relations which makes it difficult to achieve good recommendation performance.
- Trust relations are not necessarily similar across different platforms. For instance, a user following another user on Twitter may not indicate trust between them.

2.4 Inspectability and Control in Recommendation Systems

In this section two important concepts which play a key role in this master thesis, *Inspectability* and *Control* are described. Factors of Inspectability and Control have played an emerging role in areas of intelligent systems and in recent years, micro blogs. This section particularly discuss more about their impact in areas of recommendation systems.

2.4.1 Inspectability

Inspectability across recommendation systems literature is defined as the process of exposing users to the reasoning and data process behind a recommendation. In our study, we plan to incorporate inspectability in our interface which would guide the users of the system along the entire recommendation process.
Herlocker et al. in their paper [31], discuss the need to provide explanations behind a recommendation. They state that explanations provide more transparency which ultimately leads to more understanding of a recommendation. Such explanations also motivate users to incur trust in the recommendation system as they become more aware of how the system exactly works and the process behind the recommendations. Such explanations also result in providing inspectability for the recommendation system as a whole. Tintarev and Masthoff in [32],[33] study and evaluate the effectiveness of such explanations for recommender systems. They state that explanations serve seven main aims: satisfaction, transparency, effectiveness, scrutability, trust, persuasiveness and efficiency. They also conclude that explanations did indeed increase user satisfaction and also makes it easier to judge the recommendation quality.

Inspectability of the interface also increases the user's trust in the recommendation system. Bostandjiev et al. in their paper [4] designed a hybrid recommendation system which allowed users to understand and control different aspects of the recommendation process instilling factors of inspectability and control. They

concluded that explanation and inspectability increased user satisfaction and also had a positive impact on recommendation accuracy and user experience. Authors in [34] worked on an modified version of the system built in [4] where they assumed the notion of inspectability be similar to the concept of transparency stated by Tintarev and Masthoff in [35]. Their work concluded that social recommender systems and recommender systems in general do indeed can benefit from facilities that improve the inspectability. Sinha et al. in their paper [36] compare the quality of recommendations made by online systems and friends. They demonstrated that users rated systems which allowed them to have more information about the products higher compared to other systems. Ease of use and interface inspectability played critical roles to influence the users.

Much research has been done concerning recommendation systems specifically in the field of micro blogs. Inspectability and control over recommendation algorithms also provide an efficient way of dealing with vast amount of social content. Tintarev et al. in [37] focused on user experience benefits by building an inspectable and controllable interface called *HopTopics*. They stated their interface performed better than baseline interfaces in terms of perceived transparency and perceived control. In [38], the authors developed a system called *TwitInfo* for visualizing and summarizing important events on Twitter. The system also provided inspectability to the users by allowing them to manually browse through a large collection of tweets using a timeline-based display. An evaluation of the system stated that the user were efficiently able to reconstruct meaningful summaries of events in a limited amount of time. Authors in [39] built a system to support a novel interaction between a prediction algorithm and human users. The interface proved useful for the users who who not only received transparent item recommendations but also were able to understand the underlying algorithm in a better way. Authors in [40] worked on the problem of overloaded Twitter streams by building a Twitter client called *Eddi*. This interface allowed users to browse the enjoy their Twitter feed in a more efficient and enjoyable way compared to the traditional chronological interface.

In terms of other social recommendation systems, Gretarsson et al. in their work [41], developed an interactive tool which generated accurate and useful item recommendations based on Facebook profile data and user interactions with a simple graph-based interface. They concluded improved levels of satisfaction among users due to the incorporation of transparency in the system. Guy, I. et al. in [42] described a novel UI and system which provided users with novel recommendations of people who they can invite into their explicit enterprise social network. One key element of their system was the inspectability factor incorporated into their system which allowed users of the system to browse and search efficiently. Bostandjiev et al. in [43] developed an interactive visual recommendation system that combined social and semantic knowledge to produce career recommendations based on the LinkedIn API. Their experiments stated that an interactive interface allowing users factors of inspectability and control resulted better recommendation accuracy and user experience. Authors in [44] developed an interactive interface to serve the pur-

pose of explaining the recommendations in a transparent way and also to support exploration and inspectability.

In other areas of recommendation systems, Lamche et al. in their work [45] designed a system which generated interactive explanations using the current state of the user's inferred preferences. Their focus of work had two main goals: First, was to study whether a mobile recommender with interactive explanations led to more user control and transparency. Second, to develop a unique strategy to generate interactive explanations in a content-based mobile recommendation systems. The evaluation of their system demonstrated user appreciation of transparency and control in the recommendation system.

2.4.2 Control

Control can be defined as the process of allowing users to interact with different recommendation system options to tweak recommendations. Much research has already been performed on recommendation systems that support control over the recommendation process. Researchers have implemented different methods of control in their systems which range from rating items to assigning weights to item attributes. Recent research has also focused on developing interactive graphical interfaces to enable users to have control over both item and user level preferences.

Chen et al. in their paper [46] discuss the existing critiquing-based recommenders. Critiquing-based recommenders work on user's feedback during the recommendation process in order to receive better recommendations. Through the critiquing feedback, the recommender systems more accurately learn the user profiles and suggest better recommendations in the subsequent rounds. With such user based control in the recommendation systems, users are served with much better and personalized recommendations. Authors in [47] propose a novel experience-based critiquing technique to facilitate better control while making recommendations for users. Pearl Pu et al. in [48] discussed an evaluation framework to efficiently evaluate general product search and recommender systems in E-commerce environment. They stressed upon the importance of control in recommender system in order to increase recommendation accuracy as compared to baseline accuracy.

Gretarsson et al. in [41] developed *SmallWorlds*, a live Facebook application which had a interactive interface. This was used to control item predictions based upon the underlying data from the Facebook API. They also concluded that providing interactions along with visualization of the computational process provides users with better sense of control over the results. O'Donovan et al. in [49],[50] developed a collaborative recommendation system with an interactive interface allowing users to manipulate and tune different options on the interface to generate relevant recommendations. Authors in [4],[34] allow users to dynamically change and update their preferences during a recommendation process. McNee et al. in their paper [51] test user controlled interfaces against system controlled interfaces. They found user controlled interfaces provided better overall satisfaction and also

these systems better understood their tastes. Knijnenburg et al. in their research [52] studied the positive effects of user controlled interfaces on diversification. Authors in [53] evaluate the user rating behavior with preferences to rating scales for evaluating items in recommender systems.

2.4.3 Impact of Recommendation System Interfaces on Users

The main purpose that recommendation systems serve is to generate good and accurate recommendations to users. These systems have varied use in different domains such as movies, music, news articles, web pages and so on. People's opinions about items and their usability is also directly influenced by the interface of the recommendation system in use. Studies have also shown that inspectability & control has had a substantial positive impact on user experiences. In this section we look into how such recommendation system interfaces affect user's opinions.

In one of the earliest works on the topic, Cosley et al. in their work [54] perform a study about how recommendation system interfaces can influence user's opinions in two main ways: the rating scale and display of predictions at the time users rate items. The authors performed an elaborated experiment where they asked users to re-rate movies they had already rated or were unfamiliar with. The interface consisted of four main screens where the first screen had no predictions shown. The other three screens showed 10 movies with different levels of predictions with respect to user's original rating. The authors found that showing predictions when users rate movies does provide an positive impact on their ratings. Swearingen and Sinha in [55] studied the different user interactions with recommender systems to develop some general design guidelines. They discuss importance of factors such as transparency, familiar recommendations and detailed information about recommended items in the user's interaction with the system. They conclude that to design an effective interface, one must consider the following two points: one, what specific user needs are satisfied with the interaction and two, what specific systems features lead to satisfaction of those needs. They also summarize different recommendations systems in terms of inputs, rating scale, information about items and so on.

A more user-centric approach towards evaluating recommendation systems has been studied by Pearl et al. in [56]. They developed a model called *ResQue* which aims to access the perceived qualities of recommenders such as their usability, usefulness, interface and interaction qualities, user's satisfaction and the influence of the qualities on user behavior. They build on the existing TAM (Technology Acceptance Model) and SUMI (Software Usability Measurement Inventory) to make their model and identify 4 important constructs for *ResQue*: 1. user perceived qualities of the systems, 2. user beliefs of such qualities in terms of ease of use, usefulness and control, 3. subjective attitudes, and 4. behavioral intentions. Castagnos et al. in [57] discuss the on-going research in understanding the impact of decision aids on user interaction behavior. They performed an user study to understand the

influence of recommenders on how users select different items for their baskets on online stores.They concluded that user opinions about the products they want to purchase can be majorly influenced by the recommendation system. They state that the recommendation system is heavily involved in two main steps of the purchase process: helping users to refine and modify their search criteria and also to select alternatives to consider for purchase.

Jones et al. in [58] compare traditional user-controlled interfaces with more recent personalized systems using recommendations. They conducted a study to evaluate the recommendations based on implicit user feedback in comparison to users who explicitly reveal their preferences in a traditional user controlled way. Authors in [51] compared three different recommendation system interfaces to analyze the impact each interface has on user behavior. They explored the question of eliciting user preference information in the form of the following three interfaces: 1. let the system choose items for users to rate, 2. asking the users to choose items themselves, and 3. a mixed-initiative interface which combined the other two methods. The authors studied both the user-controlled and system-controlled interfaces and concluded that though both the interfaces generated high quality user models, users thought that user-controlled interfaces better understood their taste.

Schaffer et al. in their research [59], study the impact of explanations and dynamic feedback to a user during a recommendation process. Their results show improved levels of user perception of the recommendation process. The users were able to identify and remove items that contributed to poor recommendations. Their results also stated improved levels of perceived accuracy and trust in the recommender despite having lesser actual recommendation accuracy.

Pearl and Chen in their work [60],[61] discuss the factors of trust in recommendation systems. The authors work to build and develop design trust principles and algorithms for explanation interfaces. They also show significant incorporation of trust in the users as compared to traditional recommendation system interfaces. They state that users are more satisfied and perceive it to be more capable and efficient in finding better recommendations.

2.5 Evaluation

Recommendation systems form a critical part of the online community. They use opinions of the users in a given community to help individuals identify relevant content. Many different algorithmic approaches have been developed and improved in the last few years for generating effective recommendations. Each approach has their own advantages over the other. Evaluation of the recommendation systems is very crucial for a number of reasons.

- First, evaluation depicts a performance of the algorithm or the system as a whole. It states how good or efficient the recommendations are against standard baseline recommendations.

- Second, different algorithms perform differently for different datasets and systems. It is interesting to understand and decipher the performance of certain successful algorithms in proven datasets in various domains such as MovieLens & Netflix datasets.
- Third, many researchers who develop new algorithms need to compare their approach and performance against a set or existing approaches.

Shani et al. in [62] review the process of evaluation of a recommendation system. They describe three different types of experiments: offline experiments, user studies and online experiments. Performance measures such as precision, recall etc. are the easiest to perform using offline experiments by performing tests on the existing data sets. User studies are a more exhausting option where a fixed set of users are asked to perform pre-defined tasks on the system. Then, the users are asked to answer questions about their experiences. In online experiments, the system is deployed online on a large scale and the users are a part of the conducted experiments. Much of the earlier work on evaluating recommendation systems are based on how accurate the predictions are or how close are the recommendations system's predicted rating to the user's actual ratings. Recent years have seen an increasing trend in research of beyond accuracy metrics for recommendation systems [1]. In this section, two main beyond accuracy metrics namely: Novelty and Diversity are discussed.

2.5.1 Beyond Accuracy Metrics

Initially, most recommendation systems were evaluated and ranked based on their prediction accuracy. The ability to predict accurately the user's choices was mostly the basis of judging a good recommendation engine. Researchers have stated that accuracy is not always the only metric which fulfills user satisfaction [63]. Swearingen and Sinha in their work [64] show how factors apart from only accuracy can make users more satisfied. Users may also be interested in discovering novel products or in exploring more diverse items. In this sub-section, two main beyond accuracy metrics: Novelty and Diversity are discussed in detail which also play a critical role for evaluation in this master thesis.

Novelty

Novelty can be considered an important metrics for measuring user satisfaction. More researchers are inclined in the direction that novelty is one of the fundamental qualities which can be used to measure a recommendation's effectiveness. Novel recommendations are item recommendations that the user was unaware about. Zhang in [65], defines novelty as "new-original and of a kind which has not been seen before". He states that novel items should have the three following characteristics:

- Unknown: The item should be unknown to the user.

- Satisfactory: The item is satisfied for the user.
- Dissimilarity: The item is dissimilar to the items the user usually consumes.

Good novelty metrics would usually make measure how well a recommendation system was able to make a user aware of previously unknown items. The metrics would also measure to what extent the users accept the novel recommendations [63]. There exists different techniques to measure novelty in recommendation systems. One can obviously measure novelty of recommendations in a user study by asking the user how familiar they are with the item, there exist also offline measures. Two models of item novelty are discussed below: popularity based item novelty and distance based item novelty.

Popularity based novelty models are based on the assumption that less popular an item is as compared to other items, less likely it is for a user to know it. Celma in [66], define novelty for a given user as the ratio of unknown items in the list of top-N recommendations. Vargas and Castells in their work [67] [68] identify and define three ground concepts for novelty: choice, discovery and relevance. In the popularity based novelty sense, they define novelty as the difference between an item and "what has been observed". They formulate it as: *nov(i—θ)=1-p(seen—i,θ)*. In general, *p(seen—i,θ)* is a reflection of item popularity where high novelty values correspond to long-tail items that only a few users have interacted with and low novelty values correspond to popular items. They state that highly novel items can be considered as the log of the inverse polarity: $nov(i—\theta) = -log_2 p(seen—i,\theta)$.

In [67] [68] Vargas and Castells define distance based novelty as an alternative to the popularity based novelty models. They define item novelty as a distance function between the item and the context of experience. The expected or the minimum distance between the item and the set is defined as:

$$nov(i—\theta) = \sum_{j \in \theta} p(j|choose, \theta, i) d(i, j)$$

Online user based evaluations have also been performed by researchers to measure novelty [69] [56].

Diversity

Diversity is a concept well studied in the information retrieval literature. It is generally defined as the opposite of similarity. Some cases suggest that suggesting similar items is not useful for the user. Given consider an example for recommendations for a vacation, where the recommendation system recommends vacation packages to the users. A final list of 10 recommendations, all for one location, only varying on choices of hotel may not be as useful to the user as suggesting 10 different but relevant locations.

One of the most explored methods for diversity is the item-item similarity mostly based on the item content. The diversity of a recommendation list is based on the average distance between the item pairs. Ziegler et al. in [70] and Smyth et

al. in [71] were the first to approach the diversity metric in this manner. Ge et al. in their work [72] tackle the problem of placing high-diversity items in the recommendation list. They propose a guideline of organizing items in a top-N recommendation list and study how organization or placement of high-diversity items affect user satisfaction and perceived diversity.

Vargas and Castells in their work [67], state a framework for novelty and diversity on the basis of three concepts namely: choice, diversity and relevance. They propose a framework for diversity which along with the ranking also takes in account the relevance of the recommended items. Zhou et al. in [73], focus on increasing diversity with accuracy during recommendations. They work on increasing the aggregate diversity (diversity in recommendation lists for a single user) to solve the trade off between diversity and accuracy. Lathia et al. [74] work on temporal diversity by comparing different recommendation lists provided to the same user at different times. In [75], the researchers measure perceived diversity and overall attractiveness of the recommendation list.

Table 2.2 classifies the work done by researchers into perceived and inherent diversity. Perceived diversity can be defined by the diversity which is experienced by the users. Inherent diversity is the diversity which is calculated based on the dissimilarities between the recommendations on a list.

Reference	Perceived Diversity	Inherent Diversity
[70]		✓
[71]		✓
[72]	✓	
[67]		✓
[73]		✓
[74]	✓	
[75]	✓	

Table 2.2: Diversity classification.

There are two main criteria for an effective user-based search engine: *relevance* and *diversity* [76]. The quality of recommended items have become an important research are in academia and industry. Mostly, new techniques have been researched upon to improve the accuracy of the recommended items. Other aspects such as diversity of recommendations, have been overlooked. Recently, the importance of diversity with relevance has been emphasized in different studies [77] [70]. These studies state that one of the important goals of recommendation systems is to provide more personalized items to the users, and more diverse recommendations can result in more opportunities for the users to receive such personalized items in recommendations. One popular technique to achieve diversity in the recommendations is item re-ranking. Item re-ranking is basically defined as re-ordering of the items in the recommendation list based on a defined criteria. The authors in

[78] look into various ranking functions for item-based re-ranking. Some of them include re-ranking based on item popularity, item average rating, item likeability score and so on.

In order to fulfill the requirements of both *diversity* and *relevance*, different diversity-enhanced methods have been proposed. One of them is Maximal Marginal Relevance (MMR) [79]. It is one of the diversity-based re-ranking methods used for reordering ranked list of documents. Carbonell and Goldstein in [79] motivate MMR as a method to combine query-relevance with information-novelty in context of information retrieval and summarization. Authors in [80] state that it has been widely used in text summarization because of its simplicity and efficacy in nature.

2.5.2 Evaluation of Social Recommendation systems

Both traditional and social recommendation systems are most popularly evaluated by the same accuracy metrics. Ranking Accuracy and prediction accuracy are two of the most widely adopted metrics. Prediction accuracy metrics measure the closeness of predicted ratings in compared to true ratings. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are widely used by social recommender in this. Smaller values of RMSE or MAE demonstrate better performance. Ranking accuracy metrics include precision and recall measures which are the most popular. F-score is a metrics which combines both precision and recall.

2.6 Conclusions

In this chapter, different social recommendation systems along with factors of inspectability and control in recommendation system were discussed extensively. The literature involving the impact of recommendation system interfaces along with important beyond accuracy evaluation techniques have been described.

Social recommendations have lately seen an increased interest in academia and industry alike. In this chapter, the definitions of social recommendation are described and different existing social recommendation systems in various domains have been classified. Key differences between traditional recommendation systems and social recommendation systems have been discussed as well as the positive and the negative aspects about social recommendation have been described. Much of the existing research in social recommenders has focused on exploiting social relations in terms of trust and improving recommendation accuracy. Very limited work has been done in terms of recommendation interfaces involving social data to improve user satisfaction in terms of beyond accuracy metrics such as diversity, novelty etc. In terms of entertainment, movie recommendation systems are still considered to be slightly niche in the social recommendations aspect.

Literature involving factors of inspectability and control have also been discussed particularly in the areas of micro blogging. Researchers have developed interactive interfaces with micro blogging data in order to have better user experience and in turn have better recommendations and better user satisfaction. The sense of control also encourages users to return to the interface for novel recommendations. The impact involving recommendation interfaces on users based on different criteria has been evaluated as well.

In terms of evaluation metrics, two important beyond accuracy metrics novelty and diversity have been analyzed. Important characteristics of novelty and diversity have been discussed and user-centric as well as system-centric measures are classified and evaluated.

2.7 Gaps and Thesis Motivation

This section discusses the gaps found in the literature and the motivation behind the research questions and direction of this master thesis.

2.7.1 Gaps

The following is the summary of the identified gaps in the literature analyzed in this chapter:

- Very limited work has been done in terms of recommendation interfaces involving social information to improve user satisfaction.
- The study of how beyond accuracy metrics such as diversity and novelty have an impact on overall user satisfaction have not been studied extensively.
- Explainable interfaces with factors of inspectability and control especially in the field of movie recommendation systems has been explored less.
- Study of perceived diversity and perceived novelty in social recommendation systems has been limited.

2.7.2 Thesis Motivation

In this master thesis, our main motivation is to develop an interactive movie recommendation interface incorporating the following:

- Developing an interactive recommendation system interface with factors of *inspectability* and *control*.
- Leveraging social information with an existing recommendation system to improve the quality of recommendations.

- Analyzing and understanding the impact of metrics such as *system diversity*, *system novelty*, *perceived diversity* and *perceived novelty* on the *overall user satisfaction*.

Chapter 3

System

In this chapter, we look into the methodology and describe the system developed in order to answer the research questions stated in Chapter 1. Based on the literature study performed in Chapter 2, we describe and motivate all the design decisions which inspired the architecture and implementation of our system.

We begin our discussion with an overview of the system architecture. This is followed by a detailed explanation of the different components involved in the system. We describe the different phases of the system together with the underlying algorithms and datasets involved.

3.1 System Overview

In the related work, we looked into what social recommendation systems are and how they differ from traditional recommendation systems. The rapid growth of social media has encouraged the development of such systems. Social data such as social tags, tweets, social relations could be effectively incorporated in existing recommendation systems to improve the quality of recommendations. Another crucial aspect of this study, revolves around beyond accuracy metrics. Traditionally, recommendation systems have been evaluated based on the accuracy of the recommended items. Often higher accuracy does not always correspond to high user satisfaction. Other beyond accuracy metrics have been defined in an attempt to understand user satisfaction in a better manner.

In order to study our research objectives defined in Chapter 1, we focus mainly on two aspects: ***social information*** (social data such as social media tags, Facebook or Twitter followers, tweets) and ***beyond accuracy metrics*** (*Novelty* and *Diversity*).

We define the following two research goals for our system:

- ***RG1: Our goal is to design a system which can incorporate social information with an existing traditional recommendation system and recommend new and diverse items to users.***
- ***RG2: Our goal is to design a system which can help us study the relationship between beyond accuracy metrics and user satisfaction.***

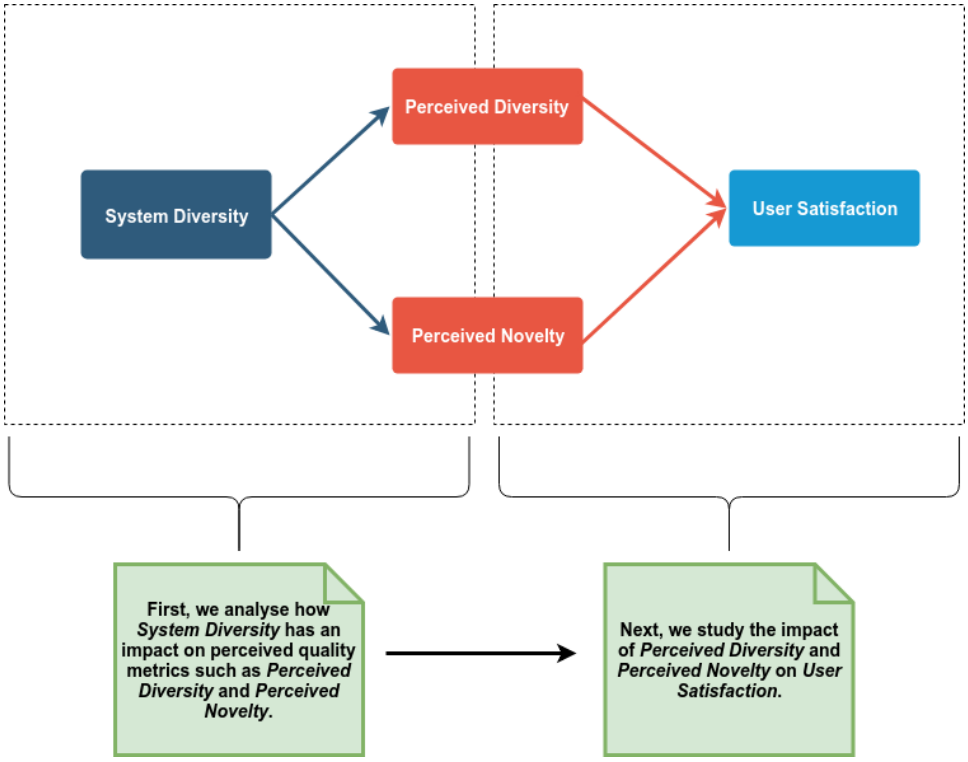


Figure 3.1: Research goals overview

Figure 3.1 gives an interpretation of our research goals defined above. We seek to measure the following:

- **Analyze the impact of *System Diversity* on *User Perception (Perceived Diversity & Perceived Novelty.)***
- **Analyze how *User Perception (Perceived Diversity & Perceived Novelty)* impacts overall *User Satisfaction.***

Based on the user-centric framework defined by Knijnenburg et al. [52], we define *system diversity* as an objective system aspect, the perceived user qualities (*perceived novelty* and *perceived diversity*) as subjective system aspects and overall *user satisfaction* as an user experience aspect for our system. Next, we first look into the architecture of the system we designed.

3.1.1 System Design

We introduce *MovieTweeters*, a web based movie recommendation interface, built to help us understand the impact of social information on the perceived quality of recommendations and to study the relationships between perceived quality of the recommended items and the overall user satisfaction.

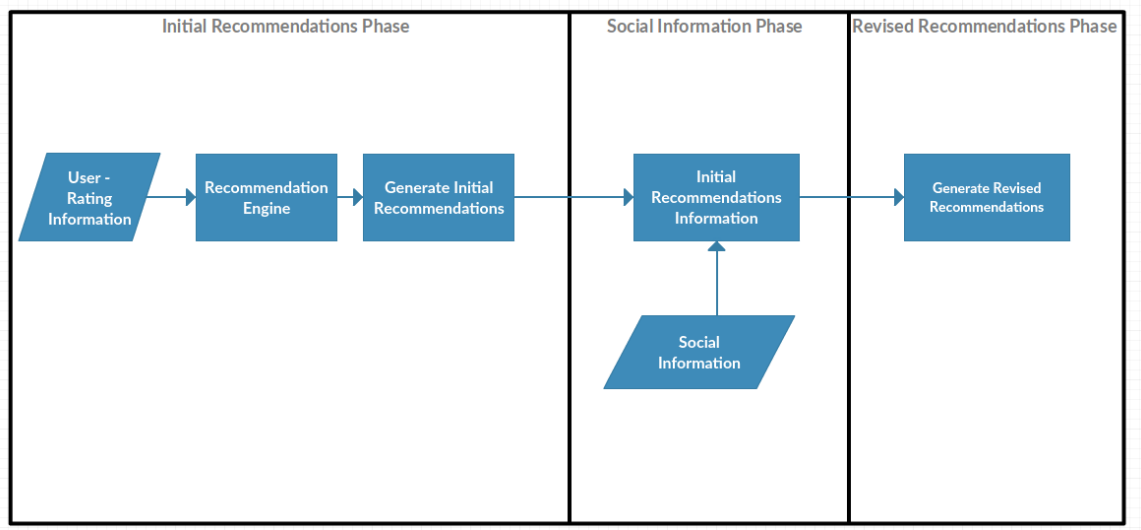


Figure 3.2: A High Level System Design of *MovieTweeters*

Figure 3.2 shows a high level overview of the *MovieTweeters* system design. The entire system is divided into three main phases namely:

- **Initial Recommendations Phase:** The initial phase involved harnessing the existing user-rating information, feeding it as an input to a recommendation engine and generating a set of initial recommendations for the user. This is the first step of the experiment which included on-boarding of new users and recommending relevant items according to their tastes. The resulting recommendations generated from this phase were used as input for the next phase of the system.
- **Social Information Phase:** The second phase of the system involved the integration of social information. The social information acted as an additional input to the system. Along with the initial recommendations, this social information was used to improve the quality of the items recommended to users.
- **Revised Recommendations Phase:** This was the final phase of the system which displayed the revised set of recommendations to the user. The recommendations were generated after incorporating the social information and initial recommendation information together.

3.2 Variable Description

We begin our analysis by first describing all the independent and dependent variables which were measured and affected during the experiment.

3.2.1 Independent Variables

There were two main independent variables in our experiment.

- **System Novelty:** *Novelty* can be considered an important metrics for measuring user satisfaction. More researchers are inclined in the direction that novelty is one of the fundamental qualities which can be used to measure a recommendation's effectiveness. Zhang in [65], defines novelty as "new-original and of a kind which has not been seen before". Following Zhang's definition, in this study we define *Novelty* as "items (movies) which the user has never consumed (seen) before."

- **System Diversity:** *Diversity* is a concept well studied in the information retrieval literature. It is generally defined as the opposite of similarity. In the context of recommender systems, diversity describes how different or dissimilar the recommended items are to the user. In the context of this study, we define *System Diversity* for the recommended items (movies) in terms of how different they are in their genres.

3.2.2 Dependent Variables

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

- **Perceived Novelty:** The extent to which users receive "new" movie recommendations. Here, we evaluate whether users are able to come across movies which they have not seen before. It is derived from the *ResQue* framework developed by Pearl et al. in [56] which accesses the quality of the recommended items.

- **Perceived Diversity:** The extent to which users felt that the recommended items were diverse to them. It is defined from the "Perceived System Qualities" stated in [56] by Pearl et al.

Perceived Novelty and *Perceived Diversity* together will be evaluated to study their relationship with *User Satisfaction*. *User Satisfaction* is defined as the user's overall experience with the system. It mainly belongs to the "Attitudes" category in the *ResQue* framework from [56].

In the upcoming sections, we will describe in detail all the underlying algorithms, approaches and datasets used in different phases of the system.

3.3 Initial Recommendations Phase

In this section, we look into the first phase of our system, *MovieTweeters*. The first phase was responsible for the following tasks:

- Using an existing movie user-rating dataset to generate initial recommendations.
- On-boarding new users in the system based on their initial tastes and movie preferences.
- Select an appropriate recommendation algorithm to build the initial recommendation engine.
- Collect user demographics of the users participating in the experiment.

3.3.1 MovieLens 1M Dataset

To on-board new users in the system and to generate the initial set of movie recommendations, we decided to use the *MovieLens 1M* dataset [81]. This is a standard evaluation data set which originally consists of 1,000,000 ratings that 6,040 users gave for 3,900 movies. As our work is primarily based in the domain of movies, the *MovieLens 1M* dataset provided us with a good starting point due to its popularity and also being a stable benchmark dataset.

Data Pre-processing

The initial step was to perform the required pre-processing on the *MovieLens 1M* dataset to make it compatible with the social information dataset, *MovieTweetings* (refer Section 3.4.1 for further details). This involved a three-step process:

1. **Dataset Coherence:** First step was to make sure that the movie titles present in the *MovieLens 1M* dataset were also present in the *MovieTweetings* dataset. Irrelevant fields such as time-stamp were removed.
2. **IMDB Details:** Second step, was to add the relevant IMDb details to the movies present in the *MovieLens 1M* dataset. The IMDb details are specified by a relevant IMDb ID. For instance, the movie Toy Story has the IMDb link http://www.imdb.com/title/tt0114709/. Here 0114709 is the IMDb ID for Toy Story. By default, the *MovieLens 1M* dataset does not contain the IMDb information for the movies. Hence, the relevant IMDb IDs for our dataset were scrapped from the bigger *MovieLens 20M* dataset.
3. **Genre Details:** The genre information of the movies play a critical role in our analysis. Hence, it was necessary to make sure all the movies listed have all the relevant genre associated with them.

Dataset	Users	Movies	Ratings
MovieLens 1M	6040	2835	964712

Table 3.1: Pre-processed MovieLens 1M dataset details

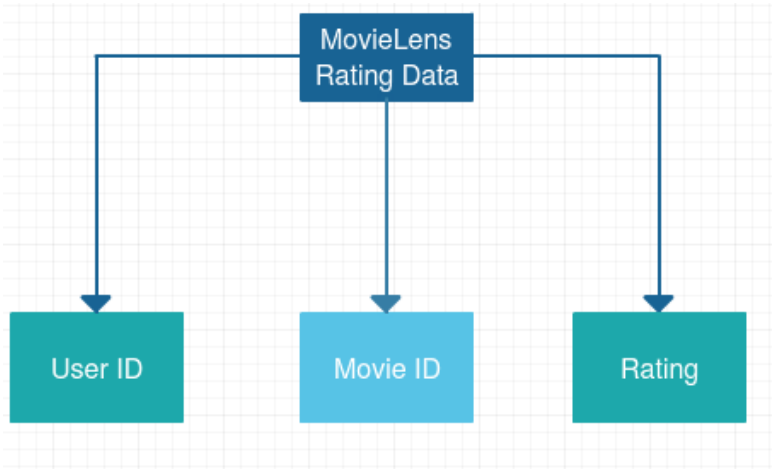


Figure 3.3: MovieLens 1M Rating data information

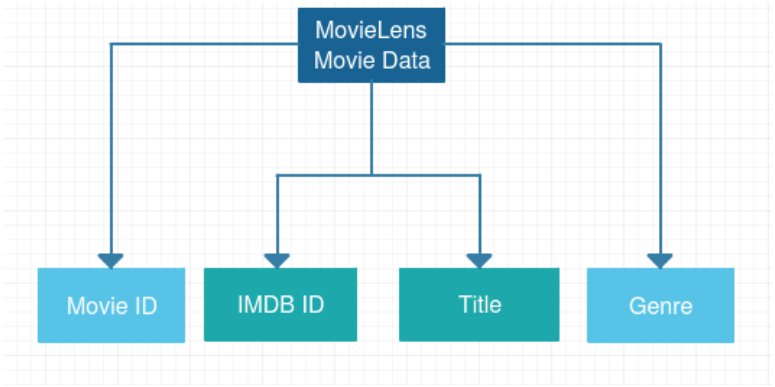


Figure 3.4: MovieLens 1M Rating data information

After performing all the necessary pre-processing steps, we had our pre-processed *MovieLens 1M* dataset which had 964712 ratings from 6040 users for 2835 movies.

The *MovieLens 1M* dataset was divided into two main data models: rating data and the movie data. Figure 3.3 and 3.4 describe the two data models in more detail.

MovieLens Rating Data Description

- **User ID:** The user ID of the MovieLens user who rated the movie.
- **Movie ID:** The MovieLens ID of the movie.
- **Rating:** The rating given to the movie by the specified user. Rating scale is from 1 to 5.

MovieLens Movie Data Description

- **Movie ID:** The MovieLens ID of the movie.
- **IMDB ID:** The IMDb ID of the movie. For instance, the movie Toy Story has the IMDb link http://www.imdb.com/title/tt0114709/. Here 0114709 is the IMDb ID for Toy Story.
- **Title:** The title of the movie.
- **Genre:** The genre description of the movie.

3.3.2 On-boarding of New Users

Recommendation systems help suggest items to users they may like based upon the knowledge about the user and the space of available items. A simple example can be given of a movie recommendation system. The movie recommendation system can keep track of the movies seen by a user who visits the system. When the user visits the system later, appropriate movies could be recommended based on the user's previous preferences. Popular recommendation algorithms such as Collaborative filtering are often used for producing relevant recommendations. Given in a domain of items, the users can express their ratings on items they have previously used. The recommendation system can compare the user's ratings to other users, find the most "similar" users or items based on a defined criterion of similarity and recommend items.

A real problem arises when new users come along and want to use the recommendation system. The system has no information about them. This *coldstart* problem is quite common in the recommendation system domain. One example of this are Collaborative Filtering systems which must acquire some information about the new user in order to make the relevant personalized recommendations. One of the most popular and direct ways to achieve this is to ask the new users to rate an initial set of items, also known as seed items.

There is one very important thing to keep in mind while performing this process of on-boarding new users. The system must be careful of selecting useful seed items that could provide valuable insight. A simple example can be of a food recommendation system. A system like that should probably not ask the new users whether they only like vanilla ice cream. Most people like vanilla ice cream, so

knowing that the new user likes vanilla ice cream gives the system very little information about the user. Due to this, the choice of initial seed items is very critical. Another factor to consider is the actual usability and interaction with the system. A good recommendation interface will reduce a new user's effort and get the user to the actual part - exploring and getting recommendations - while still obtain relevant information to make relevant recommendations. Next, we describe the method we used to select our initial seed items.

3.3.3 Initial Seed Items

Rashid, Al Mamunur, et al. in [82] discuss various strategies and the trade-offs involved while selecting an appropriate strategy to produce the initial seed items. They identify 4 dimensions that a given strategy might choose to support: (a) *User Effort* (b) *User Satisfaction* (c) *Recommendation Accuracy* (d) *System Utility*. We focused on the following 3 dimensions while we defined our initial seed strategy:

- *User Effort:* We wanted to make sure users find the sign up process convenient and easy to follow.
- *Recommendation Accuracy:* The system should be able to make relevant recommendations to the user based on their initial input.
- *System Utility:* The system should be able to serve all the users efficiently and make sure the users are exposed to good quality items when they sign up.

One such strategy is to use the *popularity* measure while determining the initial seed items. In this, the items are ranked in a decreasing order of the number of ratings. Movies are hence, presented to the user in a decreasing popularity order.

Two main issues in this strategy are listed below:

- First, this strategy will end up selecting all the popular movies from the dataset which results in a possibility that these movies are the most widely watched. If these movies are shown to the user as initial seed, little information can be garnered from that. Take for example, *Jurassic Park* is given as an initial seed item. If everyone likes *Jurassic Park* and the user using the system likes it too, the system learns little from that.

- Second, we used our pre-processed *MovieLens 1M* dataset for selecting our initial seed items. This dataset has a long tail distribution [83] of ratings for the items. This essentially means there are some movies in the dataset which have been very frequently rated (the popular movies). As seen in Figure 3.5, we can see that the items concentrated at the beginning of the graph have been most frequently rated. These are the more popular items in the data set and unpopular or new items belong to the long tail of the item distribution. Many recommendation systems would ignore such unpopular or new items and take only the most popular items in the dataset into consideration while on-boarding of new users and making recommendations. This again creates a bias where only the most popular movies are recommended.

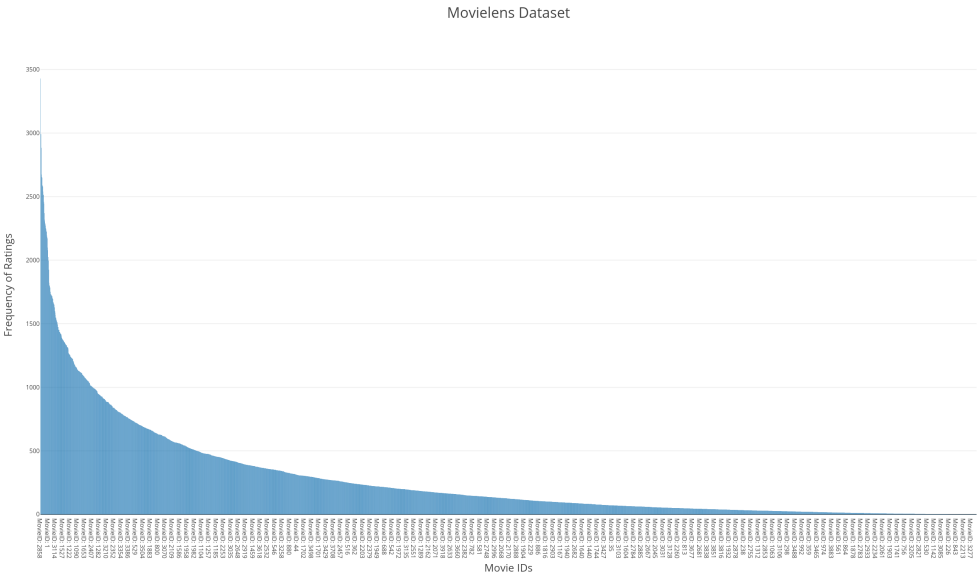


Figure 3.5: MovieLens 1M Long Tail Distribution

Our Approach

Our approach for deciding the initial seed items was based on two criteria (as defined by the authors in [82]): *popularity* and the *ratings* of the movies. We used the following approach to calculate a seed score for each item in the dataset.

$$seedscore = \log(popularity) \times \sigma^2(ratings) \tag{3.1}$$

To calculate the seed score (seen in equation 3.1), we first took the *logarithm* of the *popularity*. *Popularity* of a given movie is the frequency or the number of times the movie has been rated by users in the dataset. In the second half of the formula, we consider the *variance* of the *ratings*. This gives us a measure of how diverse the ratings have been for a particular movie. In other words, it tell us how users tend to "disagree" with the ratings of a particular movie. This helped us address the long tail problem for popular datasets as described in [83]. Our initial seed items were a combination of popular (highly rated) & unpopular movies. Keeping these two factors in mind, the initial seed items of top 40 movies (according to their seed score) were selected.

Our approach helped us achieve the following:

- Our seed items included not only popular movies (movies which were rated quite frequently) but also contained movies from the long tail of the item distribution (movies which were unpopular or rated relatively less frequently by the users). This helped us address the bias described by the researchers in [83] where only popular movies are selected for the on-boarding and recommendation process.
- This approach helped us select movies which were not only popular for different genres, but also made sure we had a mix variety of movies in the form of cult classics, foreign feature films.

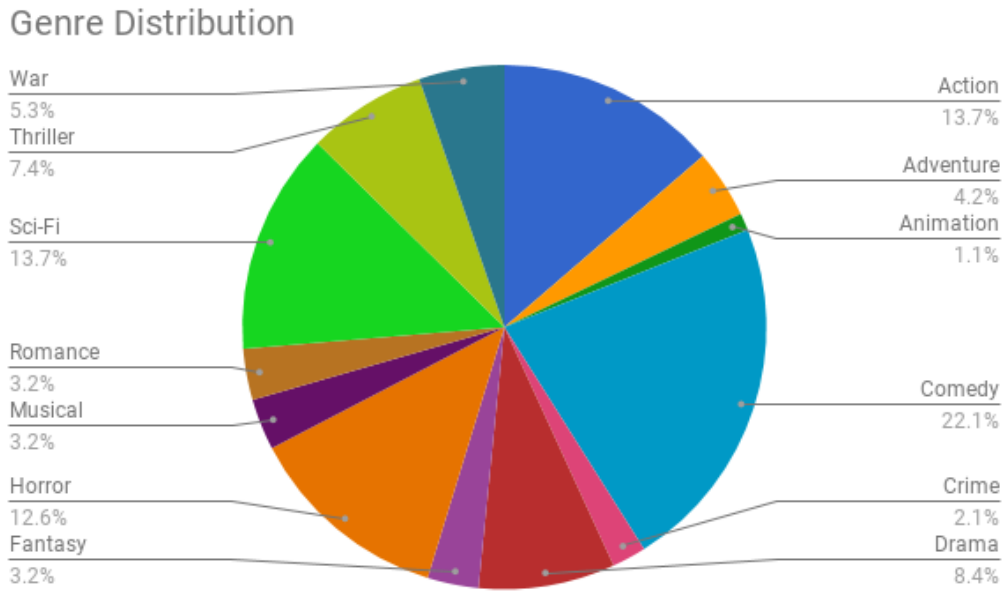


Figure 3.6: Initial Seed Items - Genre Distribution

Figure 3.6 shows the genre distribution of the movies selected for the initial seed items. The genre distribution give an idea of how varied was our selection of the movies in terms of the genres. Refer Table 3.2 to see all the movies that were selected as a part of the initial seed along with their seed score.

Movie ID	Seed Score	Movie Title
2710	5.359	Blair Witch Project, The (1999)
1924	5.080	Plan 9 from Outer Space (1958)
231	4.923	Dumb & Dumber (1994)
2657	4.909	Rocky Horror Picture Show, The (1975)
288	4.862	Natural Born Killers (1994)
2700	4.736	South Park: Bigger, Longer and Uncut (1999)
2712	4.721	Eyes Wide Shut (1999)
1676	4.438	Starship Troopers (1997)
1917	4.410	Armageddon (1998)
1721	4.347	Titanic (1997)
2427	4.290	Thin Red Line, The (1998)
1391	4.261	Mars Attacks! (1996)
2628	4.248	Star Wars: Episode I - The Phantom Menace (1999)
2459	4.248	Texas Chainsaw Massacre, The (1974)
1183	4.212	English Patient, The (1996)
2683	4.182	Austin Powers: The Spy Who Shagged Me (1999)
3785	4.176	Scary Movie (2000)
2384	4.176	Babe: Pig in the City (1998)
327	4.169	Tank Girl (1995)
3608	4.128	Pee-wee's Big Adventure (1985)
70	4.114	From Dusk Till Dawn (1996)
1573	4.070	Face/Off (1997)
2483	4.068	Day of the Beast, The (1995)
2021	4.055	Dune (1984)
3033	4.055	Spaceballs (1987)
3864	4.014	Godzilla 2000 (Gojira ni-sen mireniamu) (1999)
3156	4.007	Bicentennial Man (1999)
216	3.985	Billy Madison (1995)
780	3.975	Independence Day (ID4) (1996)
3535	3.969	American Psycho (2000)
1241	3.960	Braindead (1992)
2431	3.938	Patch Adams (1998)
2253	3.927	Toys (1992)
2976	3.927	Bringing Out the Dead (1999)
3160	3.905	Magnolia (1999)
2606	3.905	Idle Hands (1999)
2335	3.899	Waterboy, The (1998)
1088	3.876	Dirty Dancing (1987)
532	3.872	Serial Mom (1994)
3623	3.852	Mission: Impossible 2 (2000)

Table 3.2: Initial Seed Items

3.3.4 Recommendation Algorithm

One of the main purpose of a recommendation system is to apply relevant data analysis techniques to the problem at hand and help users find items that they would like to consume. The systems do this by predicting the likelihood or a list of *top-N* recommended items for a given user. These item recommendations can be generated using different techniques. Recommendations could be based on the type of items to be recommended, the demographic details of the consumers, top selling/producing items or studying the past buying/consumption behavior of the user. One of the most popular recommendation techniques is Collaborative Filtering (CF) [21] [22].

In our system, we used *Item-Item* Collaborative Filtering algorithm to generate the initial set of movie recommendations for the new users for the following reasons:

- As the main focus of this study is the incorporation of social information with an effective interface to improve the quality of recommended items, we decided to use a Collaborative Filtering algorithm to make the initial set of movie recommendations because of their popularity and simplicity of use and implementation.

- Researchers in [22] performed experiments on data from the MovieLens recommendation system and stated that *Item-based* CF techniques perform better on large datasets allowing them to scale efficiently and at the same time produce good quality recommendations.

Item-Item Collaborative Filtering Algorithm

The *Item-Item* CF approach looks into the set of items the user has rated and computes how similar they are to the target item *i*. First step, is to select *k* most similar items $\{i_1, i_2, i_3, i_4.....i_k\}$. Next, corresponding similarities $\{s_{i1}, s_{i2}, s_{i3}, s_{i4}.....s_{ik}\}$ are computed. Once the most similar items are found, the prediction is then computed by taking the weighted average of the target user's ratings on these similar items.

One of the most crucial step in the item-item CF algorithms is to compute the similarity between items and then select the most similar items. The basic idea to compute the similarity between two items *i* and *j* is to first isolate the users who have rated both these items and then apply a relevant similarity computation technique to determine the similarity $s_{i,j}$. Figure 3.7 gives a visual representation of this process. Here, the rows represent users and the columns represent items in the dataset.

In our system, we used the *Cosine-based Similarity method* to compute the similarity between 2 items. In this, the two items are represented as two vectors in the *m* dimensional user-space and the similarity between them is measured by computing the cosine of the angle between these two vectors.

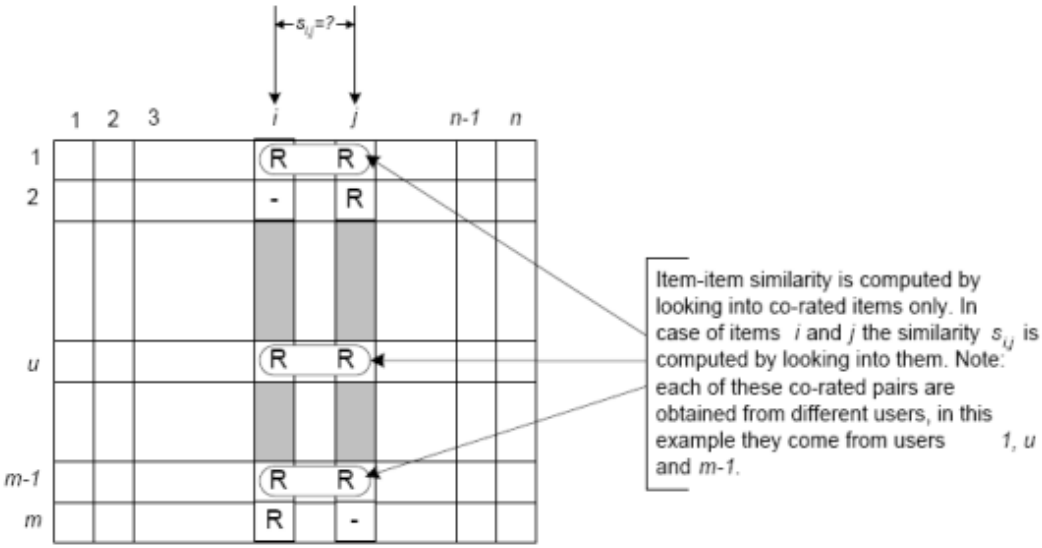


Figure 3.7: Item-based Collaborative Filtering algorithm [22]

As seen in Figure 3.7, in the $m \times n$ ratings matrix, the similarity between *i* and *j* is given by $sim(i, j)$:

$$sim(i,j) = \frac{\sum_{u \in U_{ij}} r_{ui} r_{uj}}{\sqrt{\sum_{u \in U_i} r_{ui}^2} \sqrt{\sum_{u \in U_j} r_{uj}^2}} \tag{3.2}$$

In equation 3.2, U_i is the set of users who have rated item *i*, U_j is the set of users who have rated item *j* and U_{ij} is the set of users who rated both items *i* and *j*.

3.3.5 Collecting User Demographics

Another important task of the *Initial Recommendations Phase*, was to collect the relevant user demographic information. We did this by asking the users some basic questions regarding their gender, age and their monthly movie consumption. The participants were also presented with four options where each option described a category of movie watcher. The participants were asked to select the options which described them the best. All the questions can be found in Appendix A.

3.3.6 Process Flow

The *MovieTweeters* system a web-based interface. The first step for the system user was to fill an user demographic form. Next, the system user had to rate at least 20 movies (out of 40) from the initial seed movies on a scale from 1 to 5. This scale

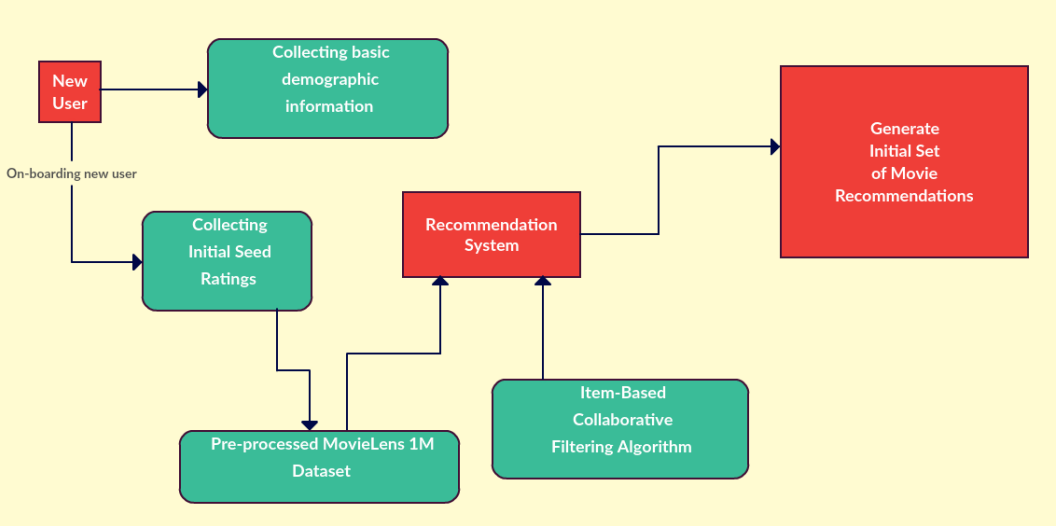


Figure 3.8: Initial Recommendations Phase Process Flow

was consistent with the rating scale (1-5) used in our pre-processed *MovieLens 1M* dataset.

Figure 3.11 shows the process flow of the first phase of our *MovieTweeters* system.

After gathering the movie preferences of the new system user, the next step was to generate a relevant initial list of movie recommendations. To perform this task, we built a recommendation system using the *GraphLab Create recommender toolkit*[1]. *The GraphLab Create recommender toolkit* provides an option to train a variety of recommender models and use them to make item recommendations. We built an item-item collaborative filtering recommender using the GraphLab recommender package. The input dataset to generate the initial recommendations was our pre-processed *MovieLens 1M* dataset which included the movie preferences (ratings) provided by the new system user of the system.A list of *top-10* movie recommendations was generated for the each system user using the system. We refer this initial list of *top-10* recommendations as **RL1**.

Following are screen shots of the *MovieTweeters* interface showcasing the above process.

[1]https://turi.com/products/create/docs/graphlab.toolkits.recommender.html

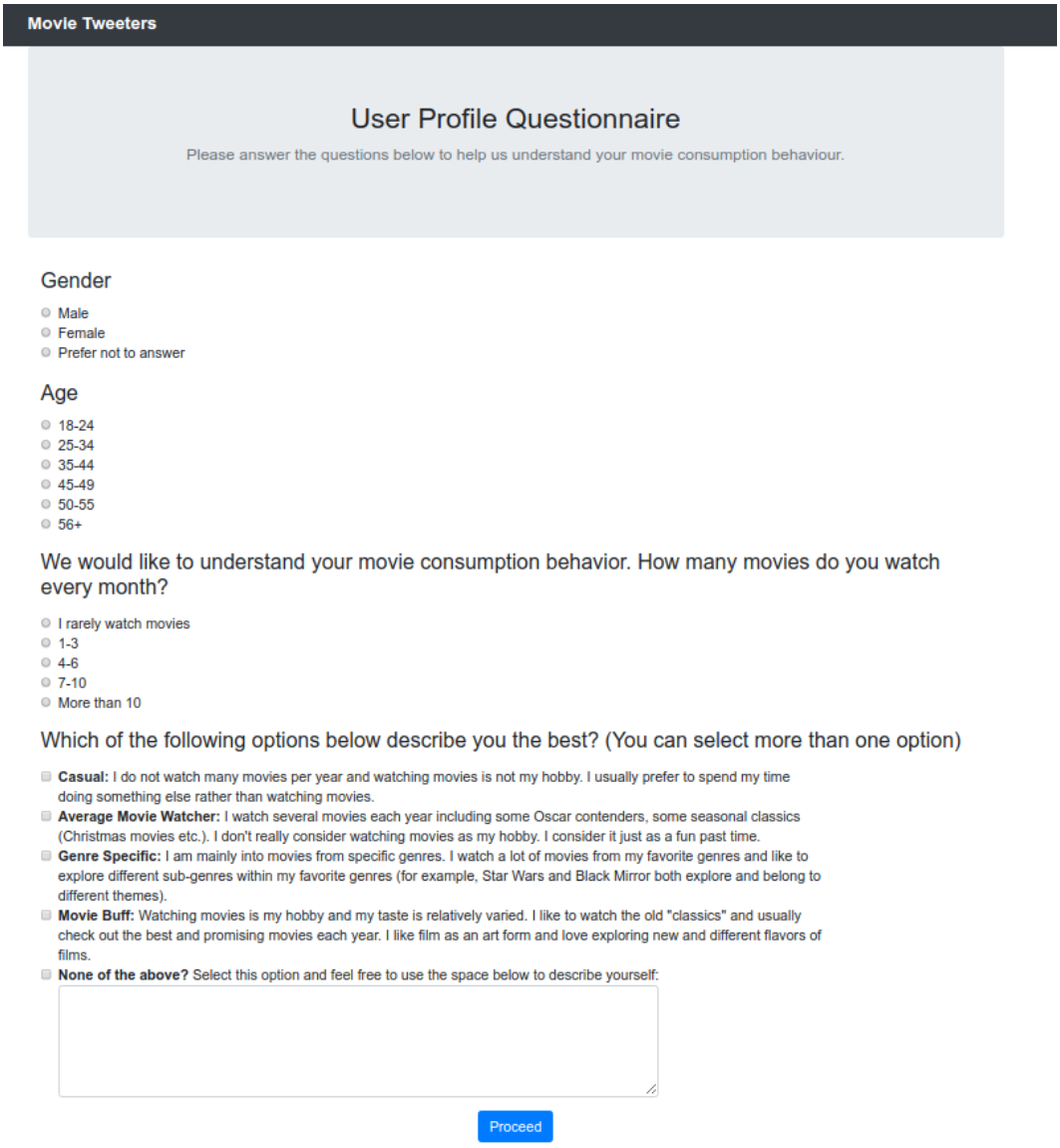


Figure 3.9: User demographics questionnaire filled by the new users

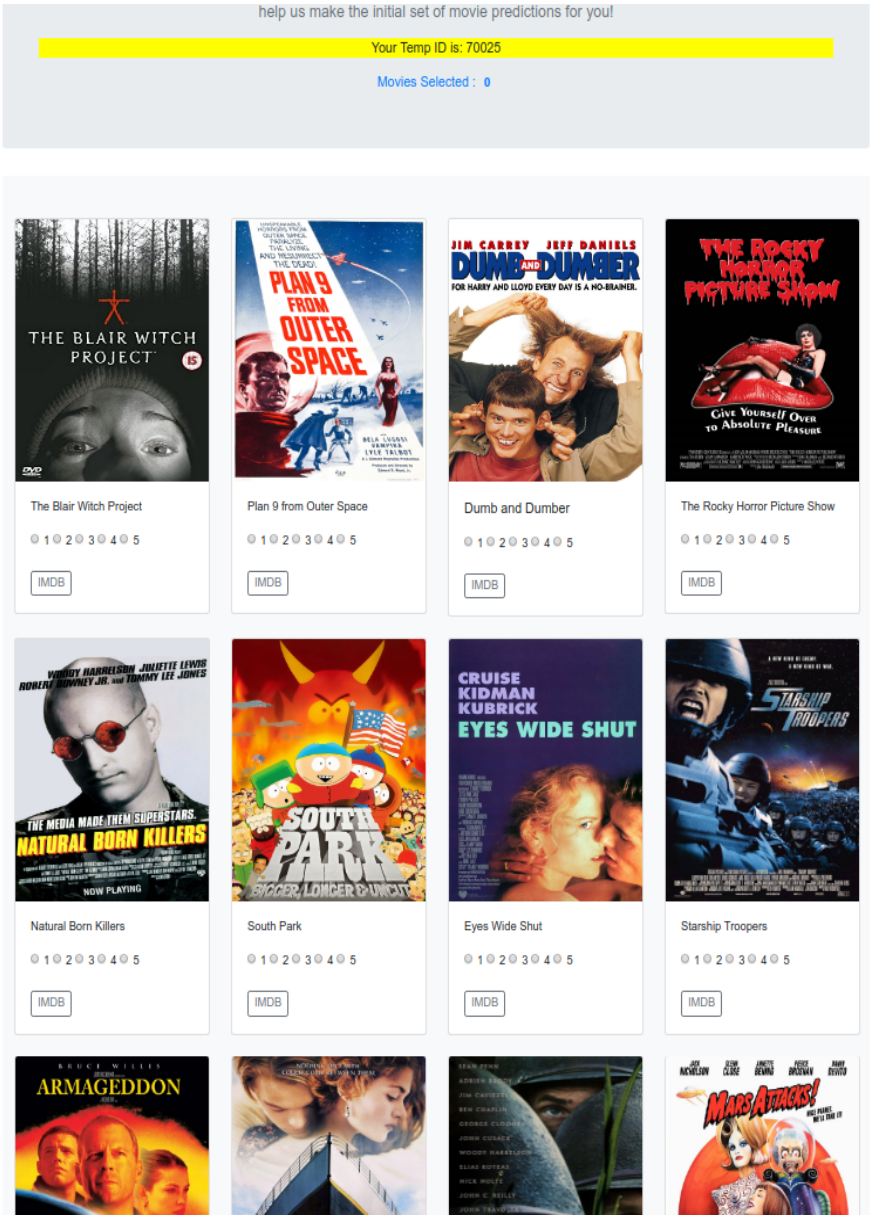


Figure 3.10: On-boarding of new users: Users were asked to rate 20 movies

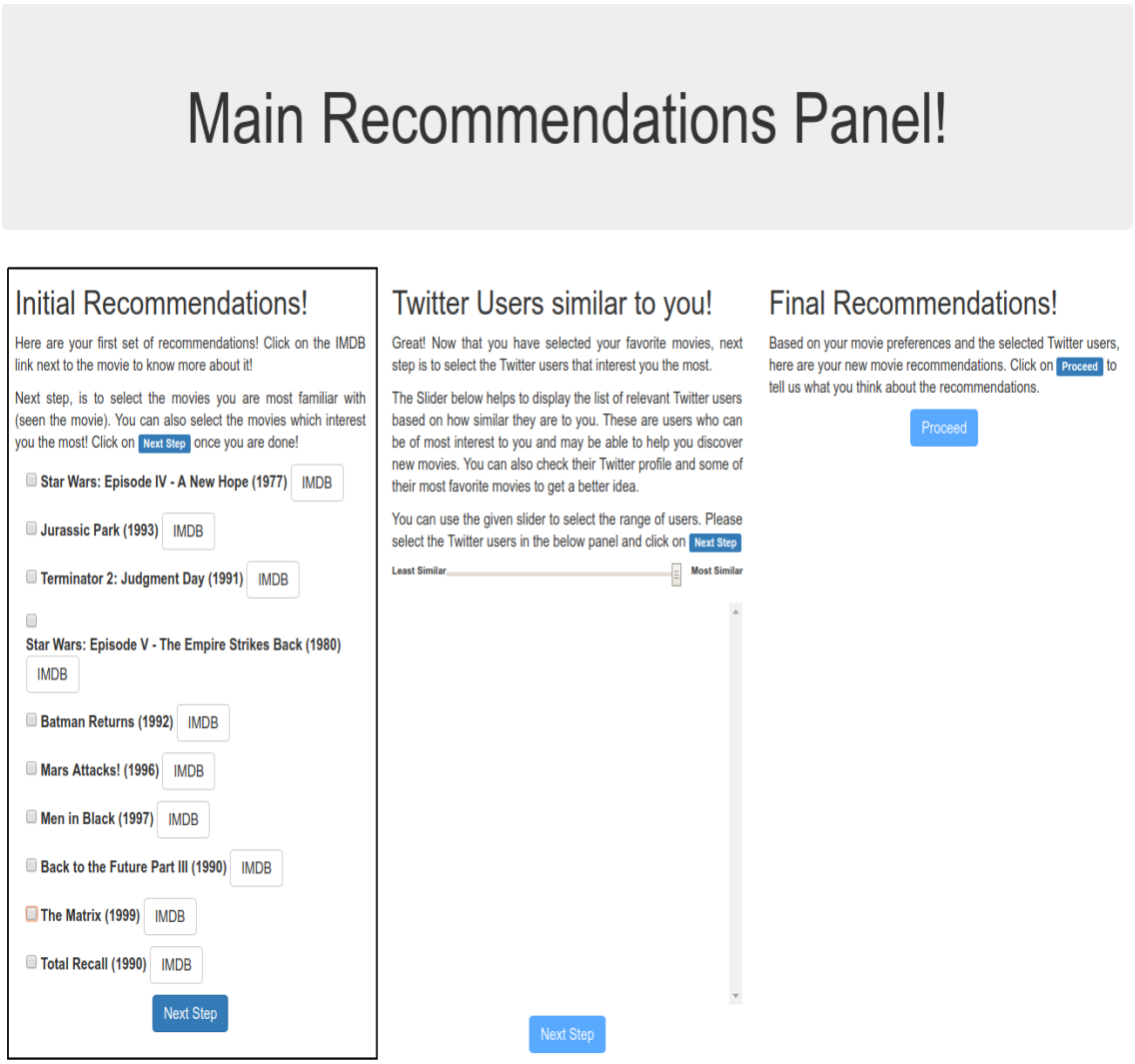


Figure 3.11: Initial Recommendations generated for the User

3.4 Social Information Phase

In this section, we look into the second phase of our system, *MovieTweeters*. The second phase was responsible for the following tasks:

- Incorporating a relevant social information dataset into our system.

Figure 3.12 shows the process flow of the second phase of our *MovieTweeters* system.

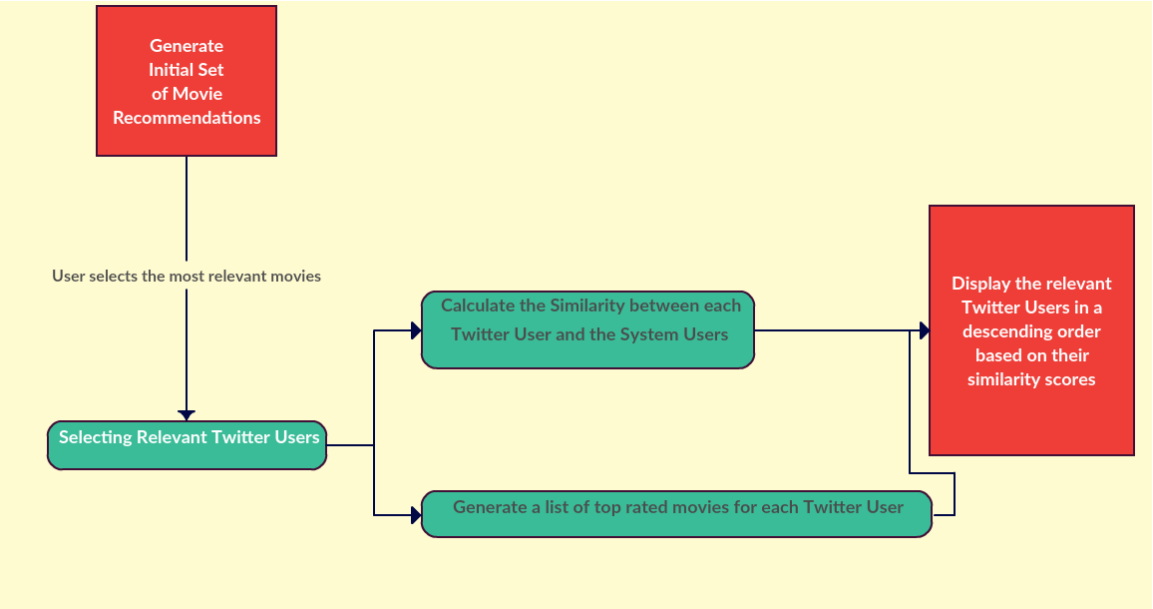


Figure 3.12: Social Information Phase Process Flow

3.4.1 MovieTweetings Dataset

We used *MovieTweetings* as our social information source. Social networks such as Facebook and Twitter have emerged as some of the most popular social media platforms allowing users to communicate and express their opinions and feedback. This online interaction between various users generates a lot of social information which could be effectively harnessed in areas such as recommendation systems. For instance, Facebook has more than 1 billion users making it one of the first social networking to do so and Twitter averaging around 360 million active users per month. The *MovieTweetings* dataset is one such example of structured social information which we use in our system *MovieTweeters* to improve the quality of our recommendations.

Item ratings form a integral part of a recommendation system as they are used to learn the user preferences. They form an indispensable component of the recommendation process. To make high quality recommendations, sufficient availability

of such ratings is of high priority. Public datasets such as MovieLens and Netflix are widely used. Especially in research, where there is a need to compare multiple recommendation algorithms and methods, such public datasets offer a very useful means for evaluation. However there are a few constraints for such public datasets:

- The datasets become gradually outdated. On such example is the *MovieLens 100K* dataset in which the most recent movie is from 1998.
- While such datasets might be useful for online evaluation, for real-time online experiments with actual users, it might not be very successful because of the lack of recent and more relevant movies.

In order to address the issues described above and also to study the impact of social information on the quality of recommendations, we used the *MovieTweetings* [84] in our system. The ratings in the *MovieTweetings* dataset originate from social media, hence they are more likely to involve more recent and relevant movies. These would serve as a valuable input data for a more user-centric evaluation experiments in the recommendation systems domain. We next describe more about the dataset in detail and the required pre-processing steps taken.

Twitter Ratings

On Twitter, tweets (short messages) can be posted and shared among users. The maximum allowed length of such tweets is up to 140 characters and user generated hashtags can also be used while making such tweets. These hashtags can be used as a unique identifier for various data extraction processes. The Internet Movie Database (IMDb) is a popular website which provides extensive information about a large number of movies. Users have the option of rating movies from 1 (lowest) to 10 (highest). This IMDb rating can also be shared on social media such as Twitter. When shared, a well structured tweet is auto generated by the system and is posted on the user's Twitter account. Some examples can be seen in Figure 3.13.

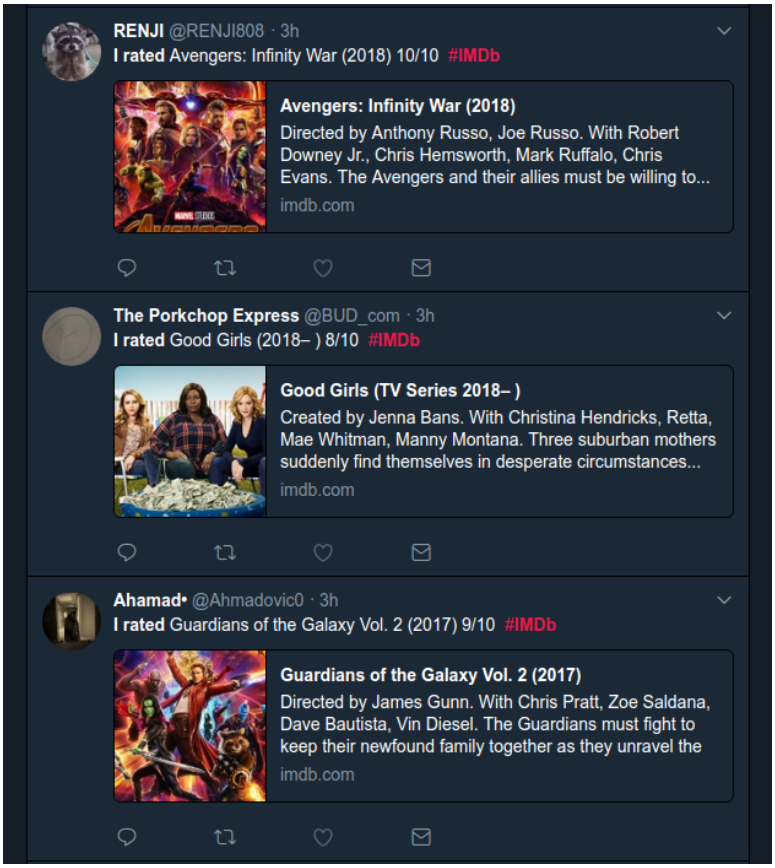


Figure 3.13: Users tweeting about their IMDb ratings

Consider the following example of one such tweet:

I rated Logan (2017) 9/10 https://www.imdb.com/title/tt3315342/ #IMDb

This tweet is well structured and can be effectively used to extract useful information. It has the name of the movie, the user-rating and also the link to the relevant IMDb page. To use this information and create a meaningful dataset, the researchers of the *MovieTweetings* dataset query the Twitter Search API on a daily basis for tweets containing the strings '*I rated*' and '*#IMDb*'. All relevant information is extracted and the appropriate IMDb pages are also cross-referenced in order to generate the relevant meta-data for the movies. The researchers have been querying the Twitter Search API daily since March 7, 2013. The dataset is still being updated daily and the most recent dataset can be found here at their GitHub[2]. For this study, we decided to use the dataset which was updated till March 2018. Next, we

[2]https://github.com/sidooms/MovieTweetings

discuss the pre-processing steps we took to make the dataset viable for our system.

Data Pre-Processing

The *MovieTweetings* dataset is aimed to be a modern version of the popular *MovieLens* datasets. Hence, similar file formats and meta-data structures were adopted. As the *MovieTweetings* dataset is updated daily, we decided to use a latest version of the dataset during the time when our system was being designed. The dataset consisted the following files:

- *ratings.dat* - Contained the user - item ratings data.
- *movies.dat* - The movie meta-data (genre- information) was present.
- *users.dat* - The unique Twitter ID of the users who rated the movies in the *ratings.dat* file.

Following were the pre-processing steps performed on the dataset:

1. **Removal of Irrelevant fields:** Irrelevant fields such as the time-stamp of the ratings were removed from the dataset.
2. **Retrieve Twitter User Screen Names:** The Twitter User IDs of the users who rated the movies were present in the dataset. We ran the Twitter API to retrieve the screen-name of these all these users.
3. **Genre Details:** It was very important that all the movies had relevant genre details. Hence, movies which did 'NA' or missing genre details were removed from the dataset.

After performing all the necessary pre-processing steps, we had our *MovieTweetings* dataset which had 606767 ratings from 45871 users for 27093 movies.

Dataset	Users	Movies	Ratings
MovieTweetings	45871	27093	606767

Table 3.3: Pre-processed MovieTweetings Dataset details

The *MovieTweetings* dataset was divided into three main data models: rating data, movie data and the user data. Figure 3.14, Figure 3.15 and Figure 3.16 describe the three data models in more detail.

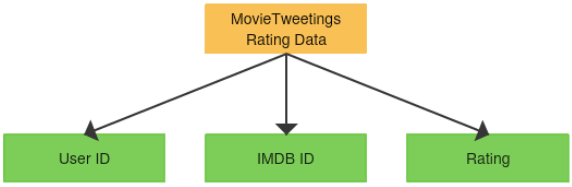


Figure 3.14: MovieTweetings Rating data information

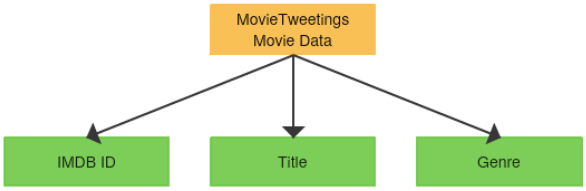


Figure 3.15: MovieTweetings Movie data information

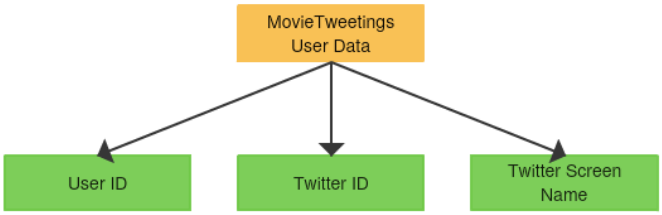


Figure 3.16: MovieTweetings User data information

MovieTweetings Rating Data Description

- **User ID:** The user ID of the *MovieTweetings* user who rated the movie.
- **IMDB ID:** The IMDb ID of the movie. For instance, the movie Toy Story has the IMDb link http://www.imdb.com/title/tt0114709/. Here 0114709 is the IMDb ID for Toy Story.
- **Rating:** The rating given to the movie by the specified user. The rating is the IMDb rating which ranges from 1 to 10.

MovieTweetings Movie Data Description

- **IMDB ID:** The IMDb ID of the movie. For instance, the movie Toy Story has the IMDB link http://www.imdb.com/title/tt0114709/. Here 0114709 is the IMDb ID for Toy Story.
- **Title:** The title of the movie.
- **Genre:** The genre description of the movie.

MovieTweetings User Data Description

- **User ID:** The user ID of the MovieTweetings user who rated the movie.
- **Twitter ID:** The Twitter ID of the MovieTweetings user. The Twitter ID is an unique value that every account on Twitter has.
- **Twitter Screen Name:** Twitter Screen Name or username of the given *MovieTweetings* user.

3.4.2 Process Flow

Figure 3.12 shows the process flow of the second phase of our *MovieTweeters* system.

The output from the first phase of the system were the *top-10* movie recommendations for the user using the system (**RL1**). The next task for the system user was to select the most *relevant movies* out of the initial recommendation list. *Relevant movies* here is defined as the movies which the system user has already watched (consumed) or the movies which seem the most interesting to him/her. This step was performed in order to get an idea of which movies in the initial recommendation list aligned with the system user's interests and consumption behavior. Figure 3.17 gives a representation of this process.

After the *relevant* movies were selected by the system user, the next step was to retrieve the relevant Twitter users from our pre-processed *MovieTweetings* dataset and display them. This was a two step process:

- **Selection of Twitter Users:** Selecting the Twitter users from our pre-processed *MovieTweetings* dataset based on *relevant* movie selection.
- **Similarity Calculation:** Calculating how similar these selected Twitter Users are to the user using the system.

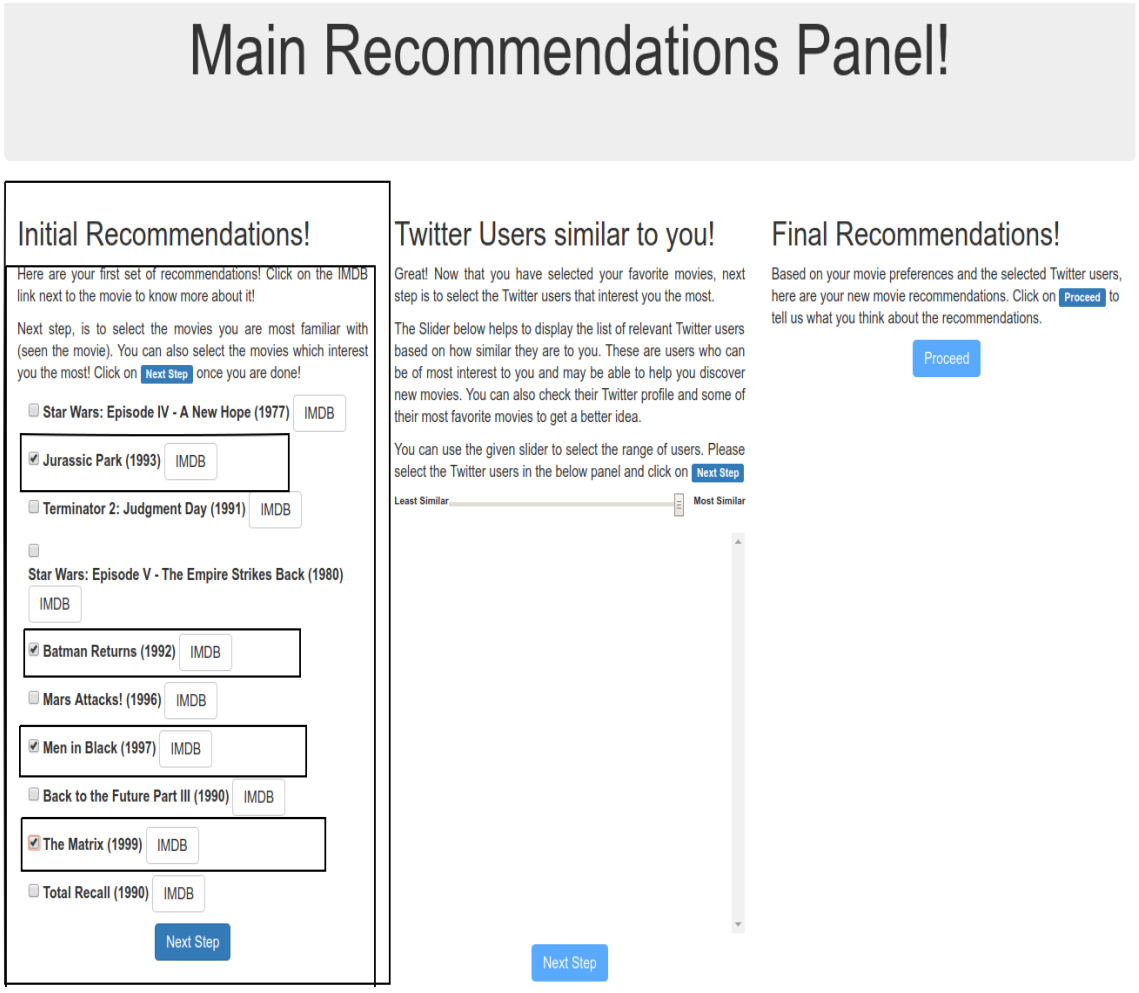


Figure 3.17: System user selecting *relevant movies*

Selection of Twitter Users:

Consider the system user selected *n* number of *relevant movies* from the initial recommendation list (**RL1**) (Figure 3.17). Next, was retrieving all the distinct Twitter Users (from the entire list of Twitter Users in our pre-processed *MovieTweetings* dataset) who had rated at least one of the *n* movies selected by the system user. Depending upon the number of movies selected (*n* movies), there could be *m* number of distinct Twitter users (3.18).

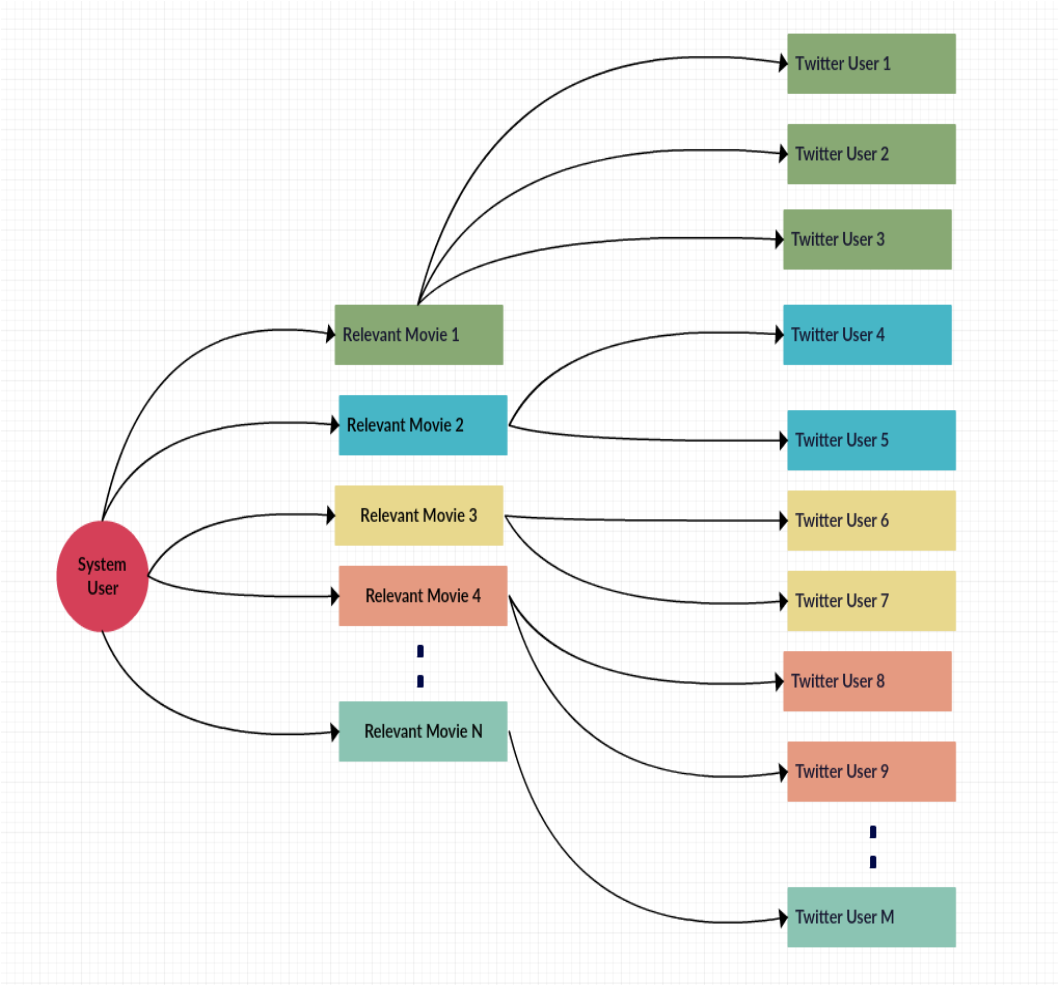


Figure 3.18: Relevant Twitter Users selection process

Similarity Calculation:

The next step involved calculating how similar the *m* Twitter users were to the system user. We used this metric to introduce ranking of these selected *m* Twitter users. We calculated this similarity by comparing the genres of the movies con-

sumed by the system user with each of the Twitter users.

A user profile of the system user was constructed based on the movies selected by the system user from the initial seed items (movies) (Section 3.3.3). This gives an indication of what genres of movies the system user had consumed.

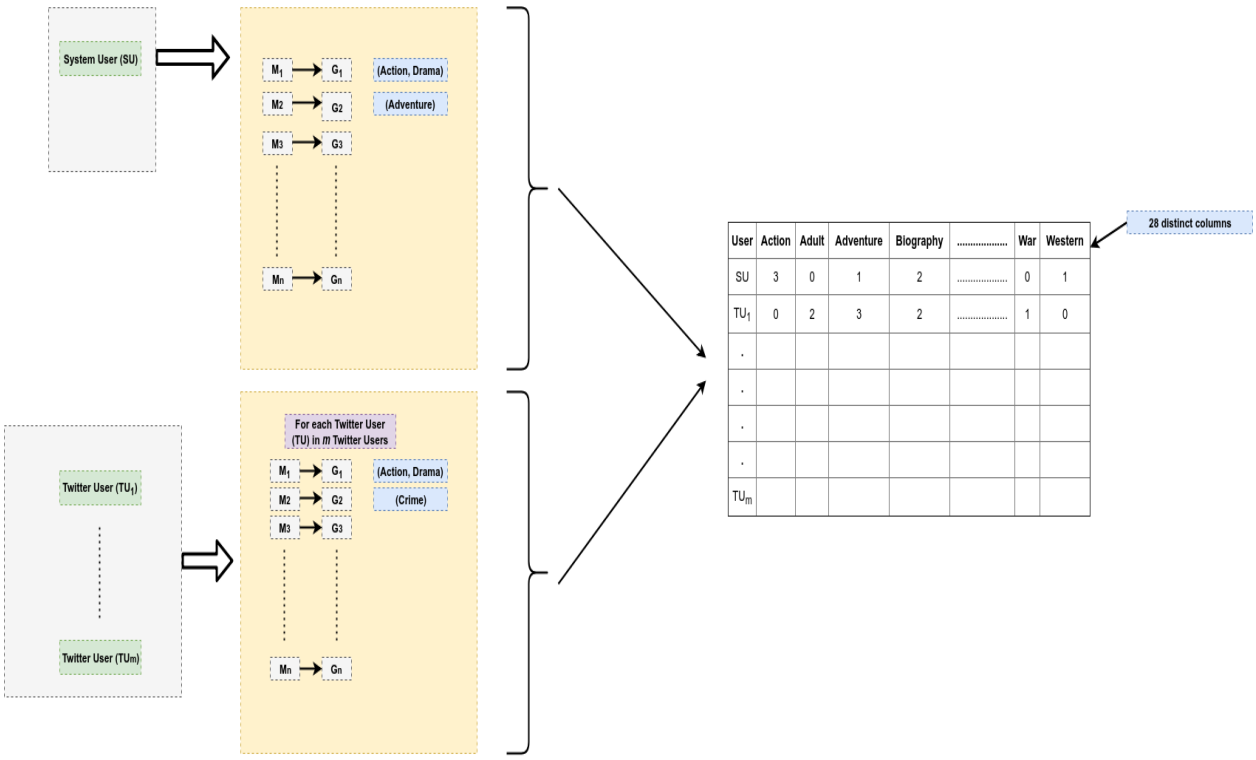


Figure 3.19: Computing similarity based on the genre consumption

Figure 3.19 gives the description of the similarity computation process. Consider a system user (**SU**) has selected *n* movies from the initial seed items (movies): $\{M_1, M_2, M_3, M_4.....M_n\}$. Each movie corresponds to a genre (or a combination of genre) given as: $\{G_1, G_2, G_3, G_4.....G_n\}$. A user-genre matrix is generated (shown in the right in Figure 3.19 where the column names consist of 28 distinct genres (identified across both the *MovieLens 1M* and *MovieTweetings* datasets). The frequency of the occurrence of each genre is counted for **SU** and the relevant count is added to the user-genre matrix.

Consider a list of *m* relevant Twitter Users (**TU**): $\{TU_1, TU_2, TU_3, TU_4.....TU_m\}$. These are the relevant Twitter users which were selected in the previous step (**Selection of Twitter Users**). Each Twitter User (**TU**), has a list of movies $\{M_1, M_2, M_3, M_4.....M_n\}$ (with their corresponding genre details $\{G_1, G_2, G_3, G_4.....G_n\}$), which they have rated in our pre-processed *MovieTweetings* dataset. For each Twitter User (**TU**), the occurrence of each genre across the movies which they have rated is counted

and the relevant count is added to the user-genre matrix.

We used the Cosine similarity (equation 3.3) to find the relevant similarity between the genre distribution of the system user (**SU**) and all the other Twitter Users (**TU**), $\{TU_1, TU_2, TU_3, TU_4.....TU_m\}$.

Given two vectors of attributes, **A** and **B**, the cosine similarity is represented using a dot product and magnitude as -

$$similarity = cos(\Theta) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_2^i} \sqrt{\sum_{i=1}^{n} B_2^i}} \tag{3.3}$$

Here, $A_i$and B_i are components of the vector **A** and **B** respectively.

For our user-genre matrix, the row which corresponds to **SU** represents the value of its vector. Similarly, the row of values corresponding to each of the Twitter Users $\{TU_1, TU_2, TU_3, TU_4.....TU_m\}$, represent the values of the corresponding vectors. We normalized the vectors and then calculated the cosine similarity of vector **SU** with all other other **TU** vectors in the matrix. This is essentially *content based filtering* as we perform the comparison between the system user and the relevant Twitter users based on the content (genre) they consume.

The final step was to arrange the Twitter Users in the decreasing order of their cosine similarity scores and display it to the system user. Figure 3.20 shows how this was achieved in our *MovieTweetings* system.

There were three main characteristics which were displayed for each of the Twitter users:

- **Twitter Profile URL:** The link to each of the Twitter profiles was shown. The system users during the experiment were also asked to learn more about the Twitter users more by exploring their Twitter profiles. This was done in order for the system users to get to know the Twitter users better.
- **Similarity Score:** An indication of how similar a given Twitter user was to the system user.
- **Most Favorite Movies:** Top favorite movies for each Twitter user were also displayed to the system users (Figure 3.21).

The system users had to select their most preferred Twitter users. A slider was also included in the system which allowed the system users to segregate Twitter users based on their similarity score. The system users had to select at least one preferred Twitter user. There was no limit on the maximum number of users selected.

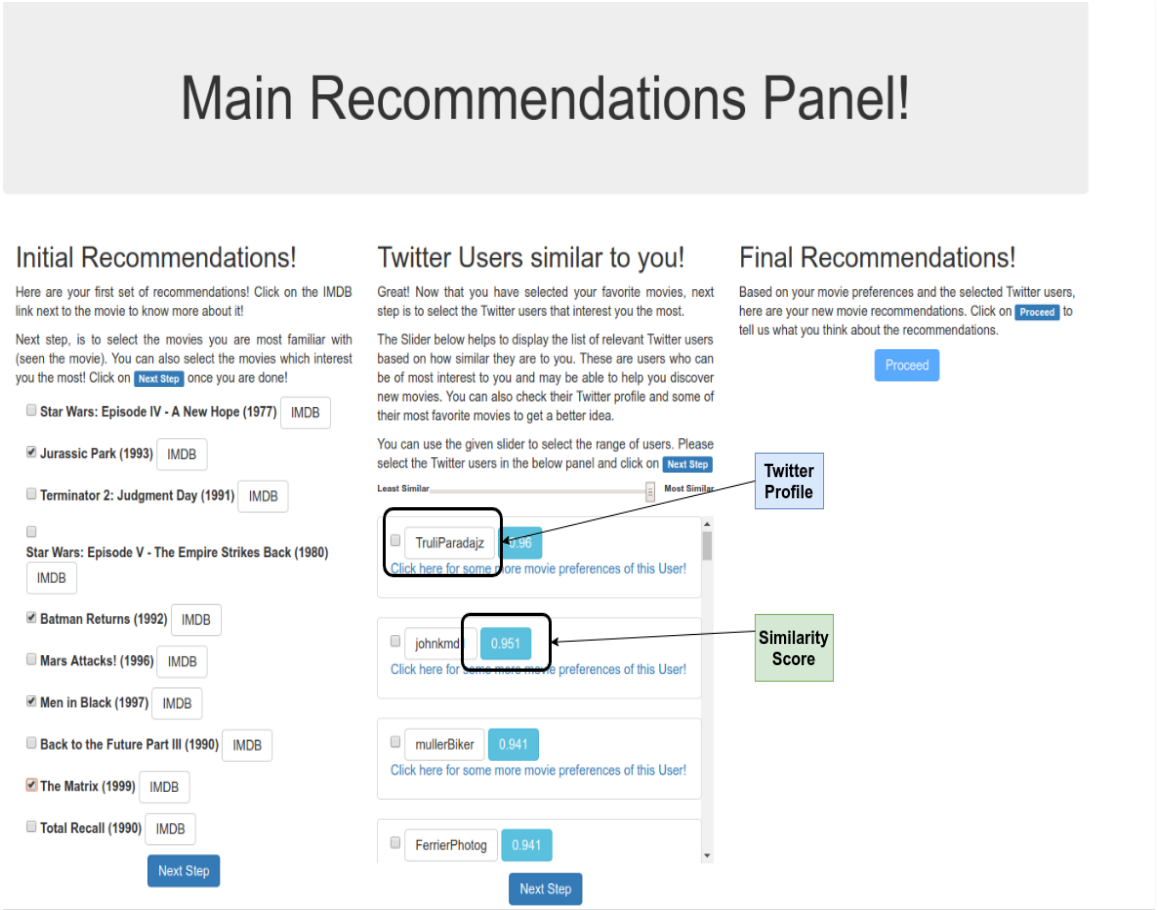


Figure 3.20: Displaying relevant Twitter Users based on their similarity scores

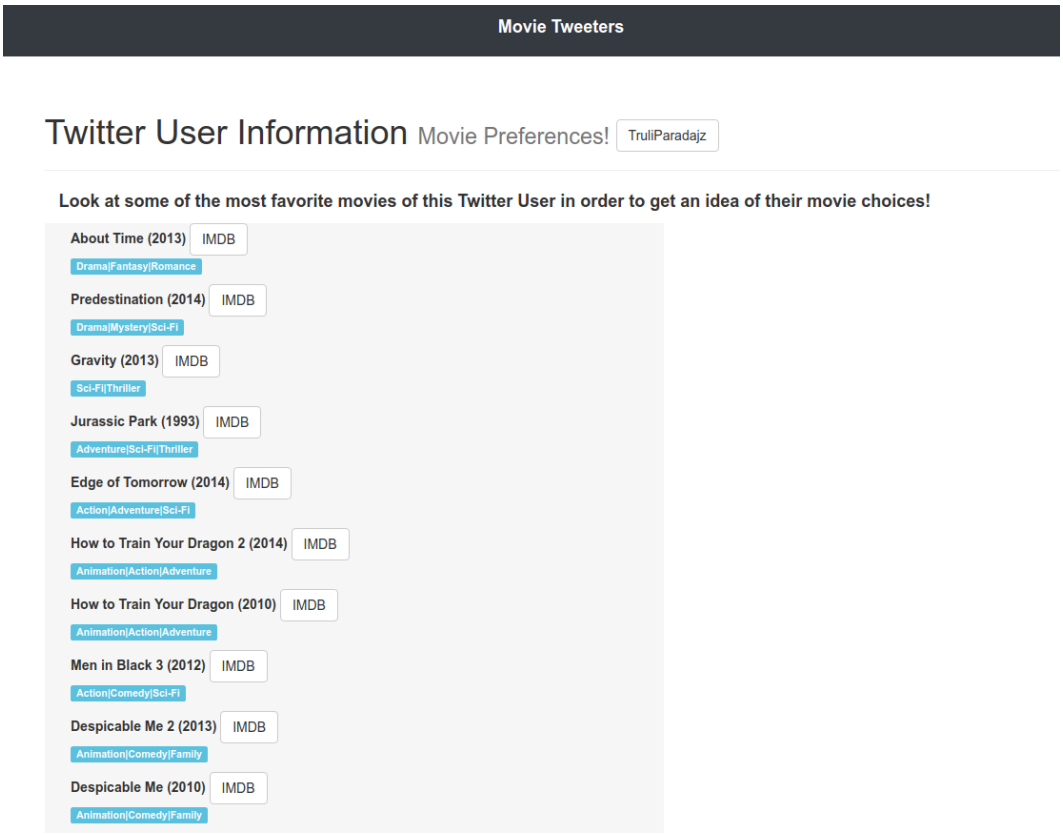


Figure 3.21: Movie preferences of a Twitter User

3.5 Revised Recommendations Phase

In this section, we look into the final phase of our system, MovieTweeters. The final phase was responsible for the following tasks:

- Retrieve the relevant list of movies based on the relevant Twitter users selected.
- Apply an appropriate re-ranking strategy to generate a revised list of movie recommendations.

3.5.1 Selection of Relevant Movies

The process is shown in 3.22 and is defined as follows:

- In the previous phase (Social Information Phase), the system user had the task to select his/her most preferred Twitter Users. See Figure 3.20.
- After the most preferred Twitter users were selected, the next step was to retrieve a list of all the movies rated by these users from our pre-processed *MovieTweetings* dataset.
- Next, we calculated the *local popularity* of each movie in the list of retrieved movies. Local *popularity* is basically the number of occurrences of the movie in the list.
- Finally, after calculating the *local popularity* of the movies in the list we created a revised movie list consisting of all the movies ranked based on their local popularity scores. We refer to this movie list as the *relevant movie list*.

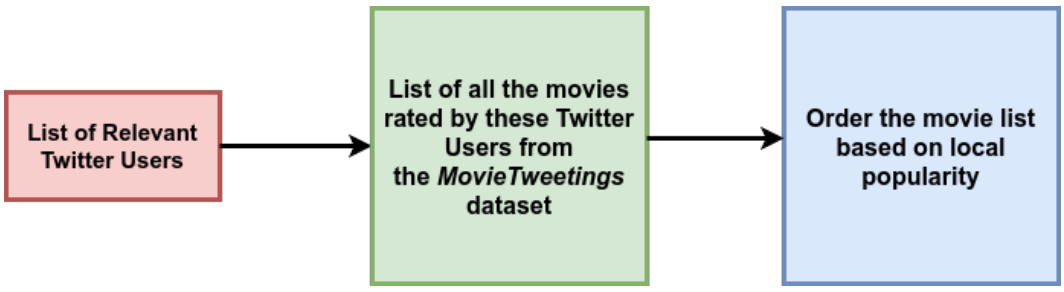


Figure 3.22: Selecting and ordering of retrieved relevant movies

3.5.2 Maximal Marginal Relevance

Following its success in the field of text retrieval and summarization, we tweak the Maximal Marginal Relevance method and apply it to our generated *relevant movie list* (Section 3.5.1). MMR for the generated *relevant movie list* was calculated using a weighted linear combination of *relevance* and *diversity* [79] [85]:

$$MMR \triangleq \max_{Di \in R/S} [\lambda(Rel(D_i, Q)) - (1 - \lambda)(\max_{D_j \in S}(1 - Sim(D_i, D_j))] \quad (3.4)$$

In equation 3.4, R refers to the ranked list of movies in the *relevant movie list*. R/S is the set difference, i.e, the set of unselected movies from R. S refers to the movies which are already retrieved.
The first half of the equation, $\lambda(Rel(D_i, Q))$ addresses the *relevance* aspect of the equation. D_i refers to a movie in the *relevant movie list*. Q for a given system user is calculated by taking into account the movies (from Initial Recommendations Phase) which were selected by him/her (Section 3.23).

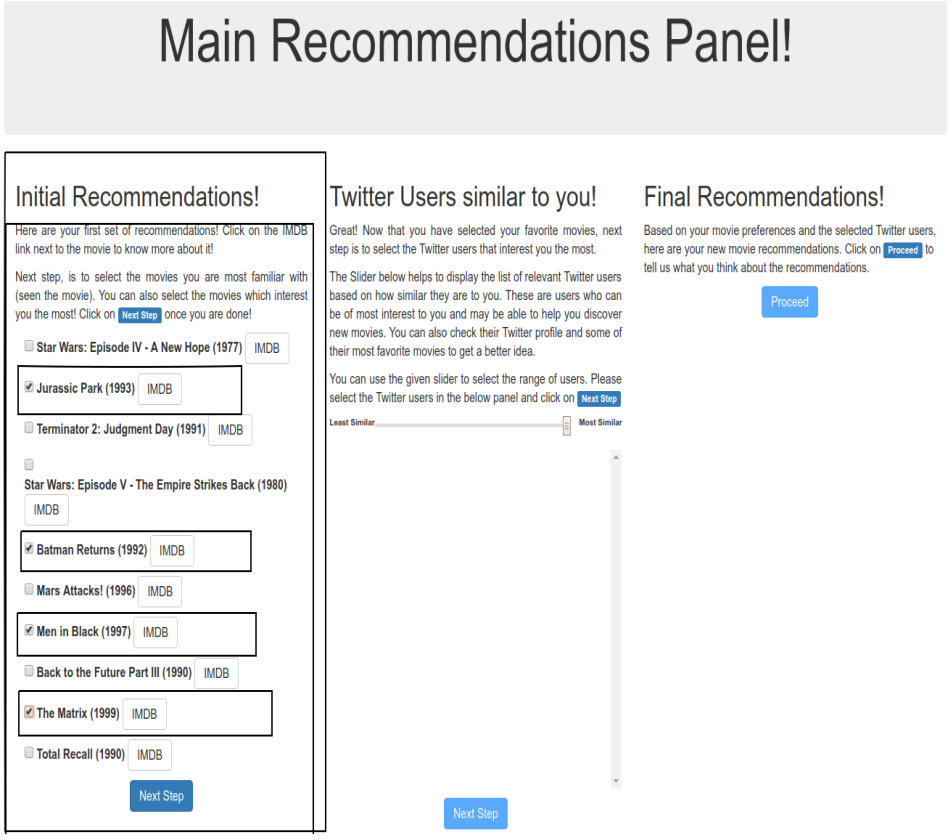


Figure 3.23: Selected movies form the Q part of the MMR equation

Here $(Rel(D_i, Q))$ is basically calculated by comparing how similar is D_i is to Q. This similarity was calculated by comparing the genres of the list of movies in Q with D_i. We used cosine similarity to achieve this and a process similar to the one described in Section 3.4.2 was performed. This gave a *relevance* score which constituted the first half of the equation.

The second half of the equation $(1 - \lambda)(\max_{D_j \in S}(1 - Sim(D_i, D_j))$ addresses the *diversity* aspect of the MMR equation. S refers to the movies which are already selected in the re-ranked MMR List. Here, cosine similarity was used to calculate how similar D_i was with D_j based on their genre details (Section 3.4.2). This constituted a score for *diversity* which formed the second half of the equation. These two scores are combined together to form the final MMR score. Based on the MMR scores, a revised list of movie recommendations, both relevant to the system user and diverse according to their tastes, was generated and presented to them. The size of the MMR list was set to be 10 (*top-10* recommendations). We refer this revised list of *top-10* recommendations as **RL2**.

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

3.5.3 Process Flow

Figure 3.24 shows the process flow of the final phase of our *MovieTweeters* system.

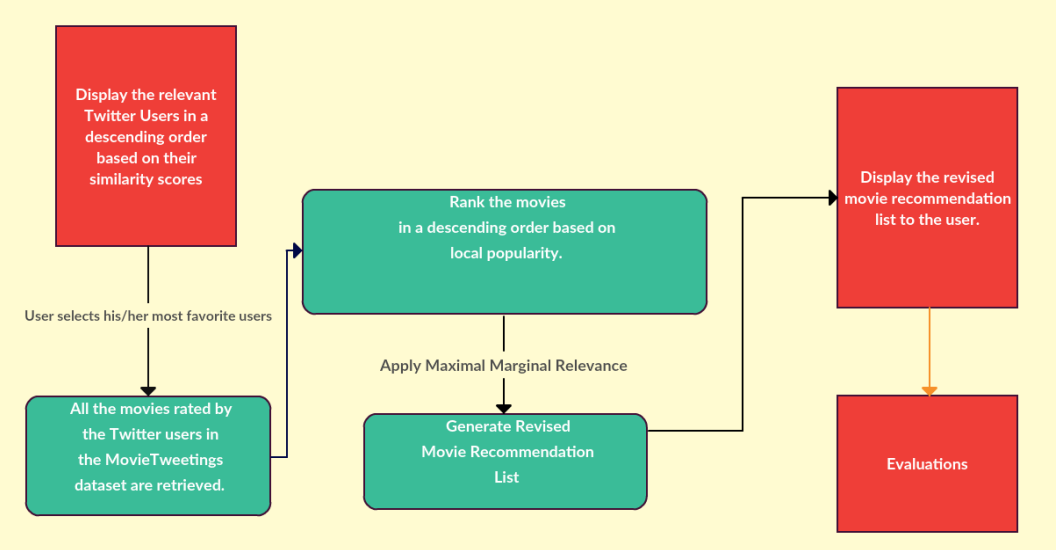


Figure 3.24: Revised Recommendations Phase Process Flow

The system users were first asked to select a list of their most favorite Twitter Users. They were asked to explore their Twitter profiles and check out their most favorite movies to get an idea about their preferences. Based on the selected Twitter Users, a list of movies from the *MovieTweetings* dataset (rated by the selected Twitter Users) were retrieved. We refer this list as the *relevant movie list*. The movies were arranged in a descending order based on the *local popularity* (Section 3.5.1) of the movies. Maximal Marginal Relevance was applied to this list and a final list of re-ranked movie recommendations was generated and presented to the user (Figure 3.25).

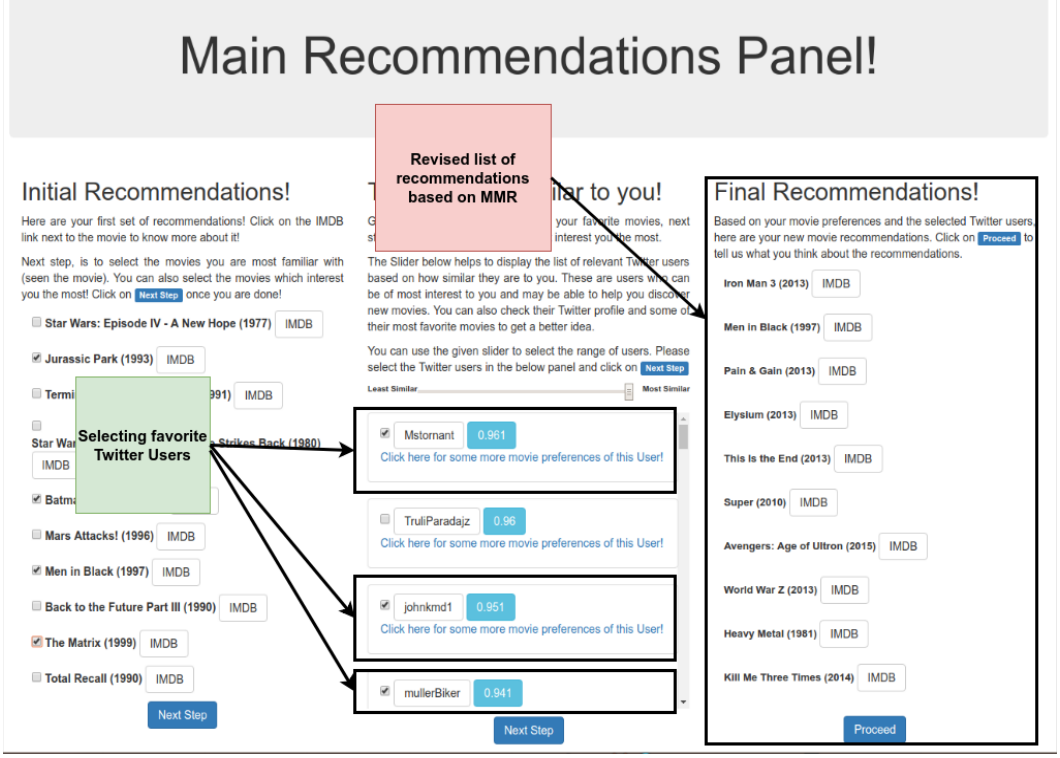


Figure 3.25: Twitter users selection and final recommendations

Chapter 4

Evaluation

We evaluate our system *MovieTweeters* in both an offline and an online setting to understand its performance and to answer the research questions defined in Chapter 1 (Section 1.2). We seek to answer the following: First, the impact of social information (with an interactive interface) on the quality of recommendations. Second, the relationship between the quality of the recommendations and user satisfaction. The offline evaluation metrics which we define later in this chapter, help us analyze how an additional input of social information with a traditional recommendation system affects factors such as *system diversity* & *system novelty*. The online evaluations help us analyze the user perceptions of the recommendations and understand their satisfaction levels.

This chapter beings with a description of the approach and the setup of the experiment. Then, we describe the evaluation metrics used to evaluate different components of the system. Finally, we discuss and analyze all the obtained results.

4.1 Approach

Our main approach was the implementation of a social recommendation system which could help study the impact of social information on the quality of recommendations. Not only that, it would also help us understand the relationship between different recommendation quality metrics with user satisfaction. To achieve this, we designed *MovieTweeters*; a web-based system which incorporated social information with an interactive interface to help users discover more novel and diverse movie recommendations (Chapter 3). To evaluate this, we designed and set up an experiment focused on helping users achieve maximum *user satisfaction* with increasing beyond accuracy factors such as *novelty* and *diversity*. Two sets of recommendation lists were generated for the users; one before the interaction with the social information and one after the interaction. We defined two main goals for the system:

- The goal was to help users get more new and diverse items in the recommendation list after the interaction. This aligns with our first research ques-

tion to observe the effects of social information on the quality of recommended items.

- Study the impact of the recommendations quality and our system on user satisfaction. This aligns with our second research question to analyze the impact of the quality of recommendations on user satisfaction.

4.1.1 Experimental Setup

In this subsection we describe the set up of the experiment.

The online evaluation was a user-based study with a limited group of participants. As our system *MovieTweeters* is a web-based system, it was hosted locally and the participants of the user study were given access to the system. The GitHub link can be found here[1]. The evaluation took place in a formal setting such as the University lab. The participants had to be over 18 years of age and it was not mandatory for them to have an existing Twitter account. The participants had to follow a sequence of pre-defined user tasks to complete the user study. The experiment was concluded with the users answering post evaluation questionnaires which were aimed at evaluating different system and perceived quality metrics.

4.1.2 Procedure

As seen in Chapter 3, our *MovieTweeters* system is divided into three main phases. Each phase had certain user tasks defined, which the underlying participants had to perform in a sequential manner. The user tasks were as follows:

- **Initial Recommendations Phase:** The first phase of the system which included on-boarding of new users and generating an initial list of movie recommendations.

    1. Participants were asked to provide basic demographic information (gender, age, movie consumption details).
    2. Participants were asked to rate at least 20 movies out of 40 movies (initial seed).

- **Social Information Phase:** The second phase of the system which involved integrating the social information dataset and generating a relevant list of Twitter users.

    1. Participants were asked to select their relevant (movies which they have already seen and/or movies which interest them the most) list of movies from the initial recommendation list (**RL1**).

[1] https://github.com/ishanghanmode/master-thesis.git

2. Participants were asked to select their most preferred Twitter users. They were asked to explore the Twitter profile and top movie preferences of the Twitter users before making any selections. They also had a slider to help them inspect the Twitter users with more ease.

- **Revised Recommendations Phase:** The phase of the system which was responsible for generating the revised list of recommendations (**RL2**).

    1. Participants were shown the revised list of recommendations based on their selections. They had to answer post evaluation questionnaires which evaluated different aspects of the recommendation lists and also their user satisfaction.

We followed a within-subjects experimental design where all the participants were exposed to both their recommendation lists (**RL1** & **RL2**) and were made to compare both the lists.

4.2 Evaluation Metrics

In this section we describe the evaluation metrics we used to measure the results obtained. We performed the evaluation with both online (with the help of the online study) and offline measures. The offline measures helped us measure the impact of social information at a system level (*system diversity* and *system novelty*). Online measures help us understand perception levels of users in terms of *Novelty*, *Diversity* and *User Satisfaction*.

4.2.1 Offline Measures

We used two offline metrics in our evaluation of *MovieTweeters*:

System Diversity:

We used *Intra List Diversity* [86] to calculate the *system diversity* of the two generated recommendation lists (**RL1** & **RL2**) in our system. It is one of the most commonly used metric in evaluating recommendation systems. It is defined as the averaged pairwise distance between all the items in a list. For our study, we define it as the following:

$$ILD = \frac{\sum_{i=1}^{n}\sum_{j=i}^{n}(1 - sim(c_i, c_j))}{n * (n-1)/2} \tag{4.1}$$

where $c_1, c_1 ... c_n$ are items in a given list and n refers to the total number of items in the list. We used the Cosine similarity measure to calculate the distance between the items.

System Novelty:

System novelty calculates how many items in the recommendation lists are new for the user (never been recommended before). We adapted the *Novelty* measure from [74] to measure the *System Novelty* of our recommendation list. There were two recommendation lists generated by our system:

- **Recommendation List 1 (RL1):** The initial recommendation list of *top-10* items (Initial Recommendations Phase). This recommendation list was generated before the interaction with the relevant social information.
- **Recommendation List 2 (RL2):** The revised recommendation list of *top-10* items (Revised Recommendations Phase). This recommendation list was generated after the interaction with the relevant social information.

The measure we used to calculate the System Novelty of **RL2** was:

$$novelty(RL2, N) = \frac{|RL2 \backslash A_t|}{N} \tag{4.2}$$

where N is the length of **RL2**. A_t refers to all the items recommended both in **RL1** and **RL2**. $|RL2 \backslash A_t|$ compares the recommended items in **RL2** with all the items in A_t.

4.2.2 Online Measures

Next, we describe the online evaluation measures we used to measure different perceived quality attributes.

Perceived Novelty

For our study, we defined *Novelty* as movies which the user has never seen before. We measured *Perceived Novelty* with the following two processes:

- **Perceived Novelty of the Recommendation Lists:** We asked the participants to select items from the two recommendation lists (**RL1** and **RL2**) which were novel for them.
- **Perceived Novelty Questionnaire:** We asked the participants to answer a set of 3 questions which helped us understand their level of *perceived novelty* across the two lists:
  - **Q 1:** The movies recommended in Recommendation List 1 were interesting to me.
  - **Q 2:** The movies recommended in Recommendation List 2 were interesting to me.
  - **Q 3:** The Twitter Users helped me obtain novel movie recommendations and improved the overall recommendation process.

Perceived Diversity

We measured *Perceived Diversity* with the following:

- **Perceived Diversity Questionnaire:** We asked the participants to answer a set of 4 questions which helped us understand their level of *perceived diversity* across the two lists:
  - **Q 1:** The list of movies in Recommendation List 2 vary from the list of movies in Recommendation List 1.
  - **Q 2:** Most of the movies in Recommendation List 2 belong to similar genres as Recommendation List 1.
  - **Q 3:** The movies recommended to me in Recommendation List 2 are diverse.
  - **Q 4:** Selecting the relevant Twitter Users helped me obtain diverse movie recommendations and improved the overall recommendation process.

User Satisfaction

We define *User Satisfaction* not only in terms of how satisfied they are with the quality of the recommendations but also their experience with the inspectability, control and overall interface aspects of our system. We measured it with the following:

- **User Satisfaction Questionnaire:** We asked the participants to answer a set of 6 questions which helped us understand their overall *user satisfaction*:
  - **Q 1:** The recommendation system provided me with good movie suggestions.
  - **Q 2:** The recommendation system helped me discover new movies.
  - **Q 3:** The movies recommended to me are diverse.
  - **Q 4:** The recommendation system made me more confident about my selection/decision.
  - **Q 5:** I am convinced I will like the movies recommended to me.
  - **Q 6:** Overall, I am satisfied with the recommendation system and the interface.

The questions were selected from the *ResQue* framework by Pearl Pu in [56]. The main motivation behind selecting the set of 6 questions was to measure the following aspects:

- Quality of the recommendations in terms of *Accuracy*, *Diversity* & *Novelty*.
- Impact of the interface on the attitude of the users.
- Overall satisfaction of the users.

4.3 Research Objectives

We designed our system *MovieTweeters*, an interactive movie recommendation system to help users discover new and diverse content. We generated two recommendation lists of movies for the users; one before **(RL1)** and one after **(RL2)** the users interaction with the social information. Our goal was to help users interact and discover their most relevant social entities (Twitter users) and eventually help them discover more novel and diverse movies in their second recommendation list. Next, we discuss our research questions and our hypotheses.

Research Question 1 (RQ1)

Does **incorporating social information** with an **interactive interface** help users **discover novel and diverse items**?

In order to answer **RQ1**, we analyzed the impact of *System Diversity* on *Perceived Diversity* and *Perceived Novelty*. We also study the change in *Perceived Novelty* of the recommended items from both the recommendation lists (**RL1** and **RL2**) for the participants.

We form the following hypotheses for answering **RQ1**:

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

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

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

Research Question 2 (RQ2)

Does **discovering novel and diverse items** help achieve **better levels of user satisfaction**?

In order to answer **RQ2**, we analyze the impact of *Perceived Novelty* and *Perceived Diversity* on *User Satisfaction*.

We form the following hypotheses for answering **RQ2**:

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

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

4.4 Results

In this section, we discuss in detail the obtained results. We describe in detail about the participants involved in the user study, the data collected, our evaluations of the data and finally attempt to answer our research objectives of finding the impact of our interactive interface (with relevant social information) on the system and perceived quality of recommendations and overall satisfaction.

4.4.1 Participants

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

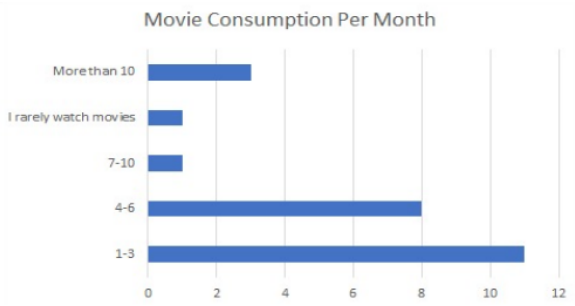


Figure 4.1: Movie consumption behavior

Figure 4.1 gives us an overview of their movie consumption patterns with most participants in the range of 1-6 movie consumption per month.

4.4.2 Data Collection

Next, we look into how the data was collected from the participants. We describe the user questionnaires asked to evaluate the perceived quality attributes (*Perceived Novelty* and *Perceived Diversity*) and the overall *User Satisfaction*.

Perceived Novelty

For understanding the *Perceived Novelty*, we collected the data in two forms:

- **Perceived Novelty of the Recommendation Lists:** We displayed the two generated recommendation lists (**RL1** & **RL2**) for the participants and asked them to select movies which were novel for them (example shown in Figure 4.2).

- **Perceived Novelty Questionnaire:** We asked the 3 questions (Q1 - Q3) defined for *Perceived Novelty* in Section 4.2.2.

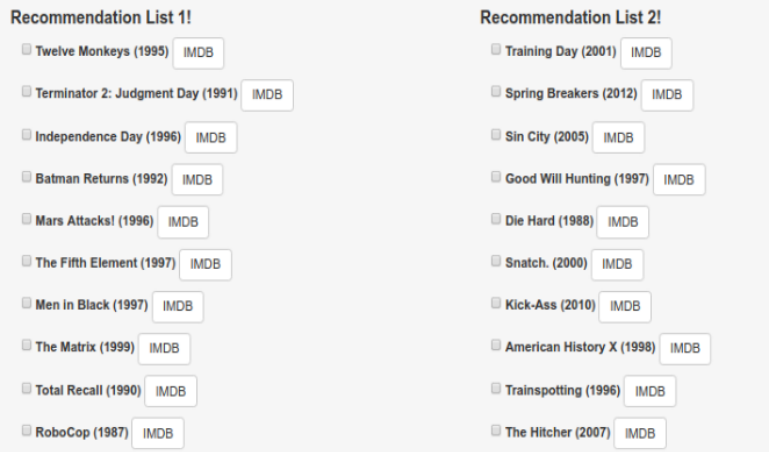


Figure 4.2: Perceived Novelty of the Recommendation Lists

Appendix B (Table B.1) contains the data collected for all the 23 participants. The overall score was calculated as an average of all the response to Q1, Q2 and Q3. All responses to the questions (Q1 to Q3) were in the form of a Likert scale (1 - 5). The overall score distribution of the participants is shown in Figure 4.3

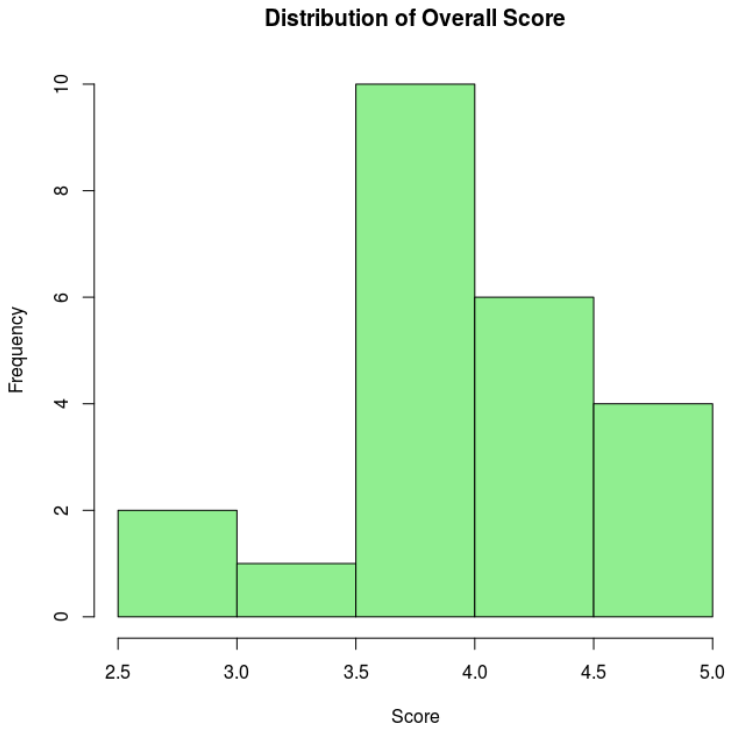


Figure 4.3: Perceived Novelty - Overall Score Distribution

Perceived Diversity

For understanding the *Perceived Diversity*, we collected the data in the following way:

- **Perceived Diversity Questionnaire:** We asked the 4 questions (Q1 - Q4) defined in Section 4.2.2 to understand the *Perceived Diversity* of the participants.

Appendix B (Table B.2) contains the data collected for all the 23 participants. The overall score was calculated as an average of all the response to Q1, Q2, Q3 and Q4. All responses to the questions (Q1 to Q4) were in the form of a Likert scale (1 - 5). The overall score distribution of the participants is shown in Figure 4.4

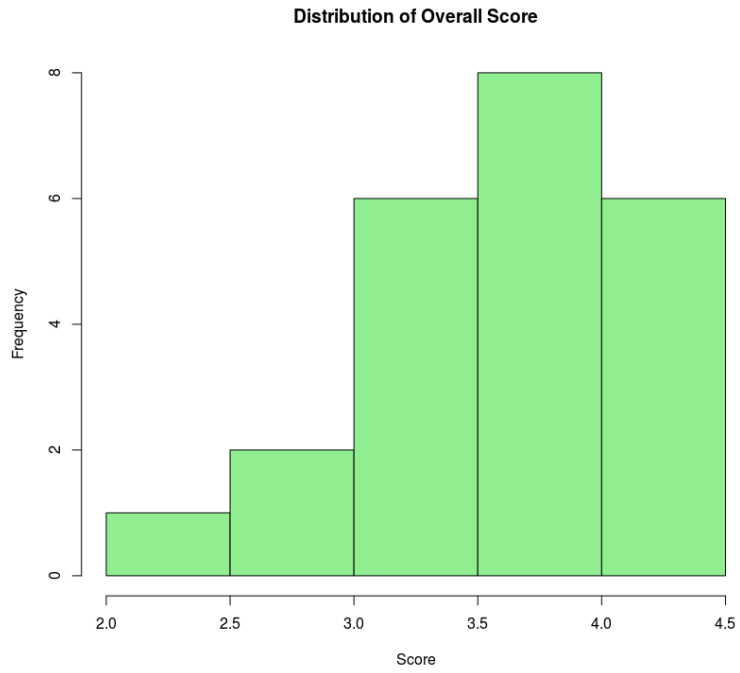


Figure 4.4: Perceived Diversity - Overall Score Distribution

User Satisfaction

For understanding the *User Satisfaction*, we collected the data in the following way:

- **User Satisfaction Questionnaire:** We asked the 6 questions defined in Section 4.2.2 to understand the *User Satisfaction* of the participants in terms of the recommended items and the interface as a whole.

Appendix B (Table B.3) shows the data collected for all the 23 participants. The overall score was calculated as an average of all the response to Q1, Q2, Q3, Q4, Q5 and Q6. All responses to the questions (Q1 to Q6) were in the form of a Likert scale (1 - 5). The overall score distribution of the participants is shown in Figure 4.5.

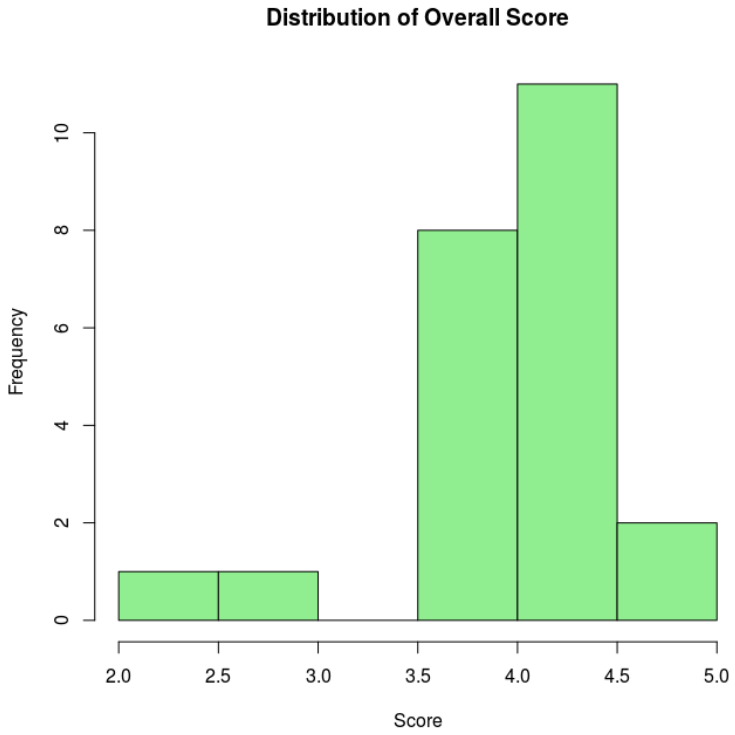


Figure 4.5: User Satisfaction - Overall Score Distribution

System Novelty

Using the measure described in Section 4.2.1, we calculated the *System Novelty* of the recommendation list 2 (**RL2**). We did this to analyze how the number of novel items were affected after the participants interaction with the social information. Appendix B (Table B.4) shows the System Novelty scores achieved for each participant. The system novelty distribution of the participants is shown in Figure 4.6.

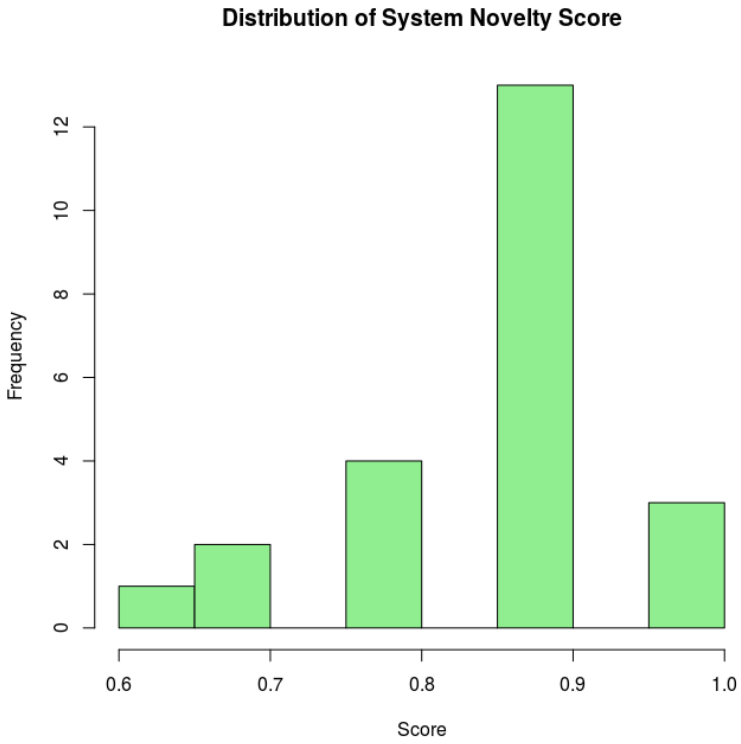


Figure 4.6: System Novelty Score Distribution

System Diversity

Using the measure described in Section 4.2.1 , we calculated the *System Diversity* of both the recommendation lists (**RL1** & **RL2**). Overall, we found a significant increase in the system diversity across the two recommendation lists (Figure 4.7).

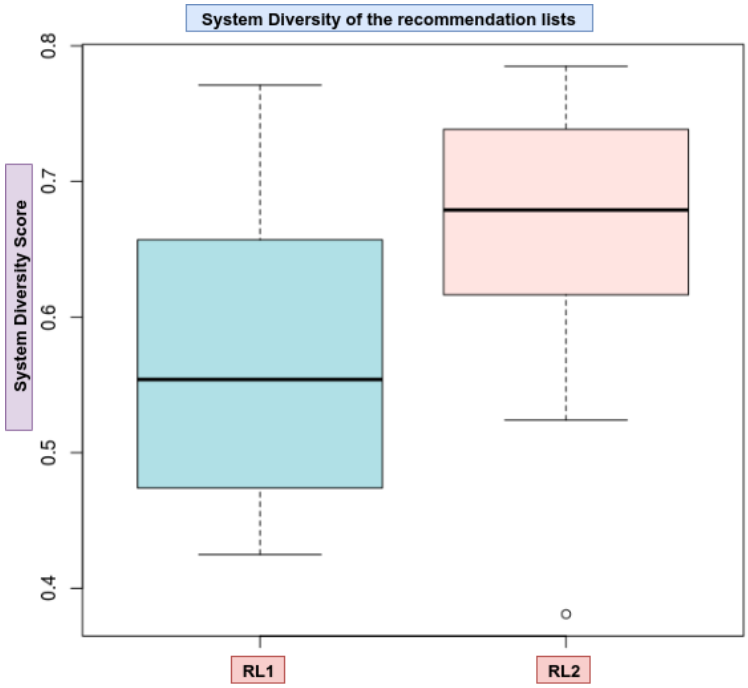


Figure 4.7: System Diversity Scores, before (RL1) and after (RL2) the interaction.

The Δ score gave us an idea of how the system diversity changed between the two recommendation lists.

The Δ score was calculated using the following formula:

$$\Delta Score = |RL2 - RL1| \tag{4.3}$$

Appendix B (Table B.5) shows the System Diversity scores achieved for each participant. The system diversity distribution of the participants is shown in Figure 4.8.

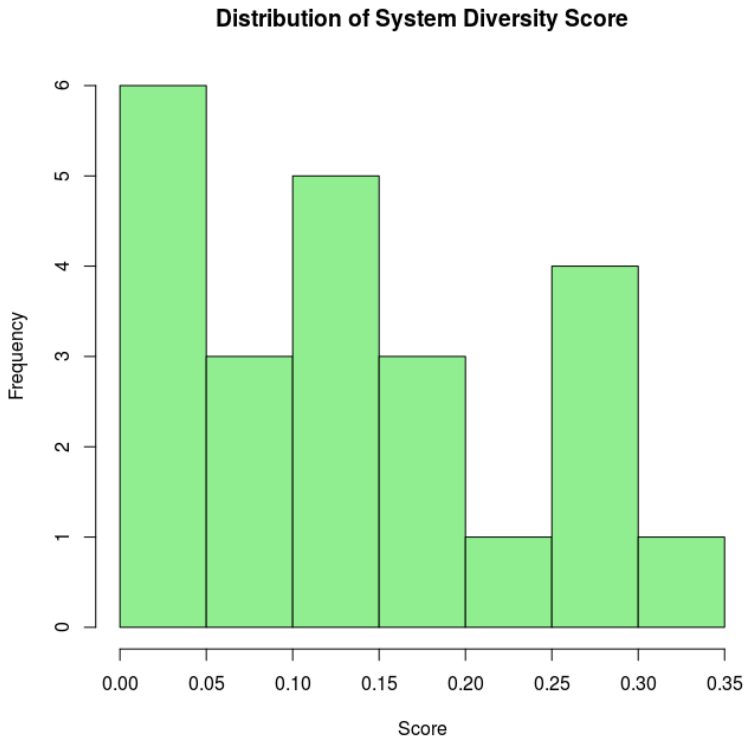


Figure 4.8: System Diversity Score Distribution

4.4.3 Hypotheses Results

We designed our system *MovieTweeters*, an interactive movie recommendation system. Our goal was to help users interact and discover their most relevant social entities (Twitter users) and eventually help them discover more novel and diverse movies in their second recommendation list.

Next, we discuss our research questions and our hypotheses.

Research Question 1 (RQ1)

Does **incorporating social information** with an **interactive interface** help users **discover novel and diverse items**?

We have the following hypotheses for answering **RQ1**:

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

To validate our Hypothesis H_1, we ran the Spearman's Rank-Order Correlation test to compare the impact of the $\Delta Score$ of *System Diversity* on the *Perceived Diversity* of the participants on the recommended items from both the recommendation lists (**RL1** and **RL2**). We used the responses to the following question to determine the *Perceived Diversity*:

- *The list of movies in Recommendation List 2 vary from the list of movies in Recommendation List 1.*

We made this selection based on the fact that the more than half of the participants responded about comparing the movies for this particular question in terms of different genres, type of movies (cult classics/popular/new/old) etc.

Table 4.1 gives more information about the data statistics (min, max, mean, standard deviation values).

Data Type	**Min.**	**Mean**	**Max.**	**Std D.**
Δ Score (System Diversity)	0.0020	0.14	0.31	0.09
Perceived Diversity	3.00	4.17	5.00	0.57

Table 4.1: System Diversity & Perceived Diversity Data Statistics

Test	**p-value**	**rho**
Spearmans rank-order correlation	0.007	-0.54

Table 4.2: System Diversity & Perceived Diversity Correlation

As seen from Table 4.2, the ρ **value** is **-0.54** and it is significant ($\mathbf{p = 0.007}, \mathbf{p < 0.05}$). This rejects the null hypothesis and our alternative hypothesis (H_1) is accepted. The ρ value shows that there is a strong negative correlation between the

System Diversity and the *Perceived Diversity* of the recommended items in both the recommendation lists.

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

To validate our Hypothesis H_2, we ran the Spearman's Rank-Order Correlation test to compare the impact of the $\Delta Score$ of *System Diversity* on the *Perceived Novelty* of the participants on the recommended items from both the recommendation lists (**RL1** and **RL2**). We used the responses to the following question to determine the *Perceived Novelty*:

- *The Twitter Users helped me obtain novel movie recommendations and improved the overall recommendation process.*

As we aim to find out what impact the difference in the *system diversity* has on the discovery perception, the above question gives a more accurate idea of how users perceived *Novelty*. The comments by the participants after the experiment also suggested the same, as the feedback given by the them was about how they analyzed novel movies from both the recommendation lists (**RL1** and **RL2**).

Table 4.3 gives more information about the data statistics (min, max, mean, standard deviation values).

Data Type	**Min.**	**Mean**	**Max.**	**Std D.**
Δ Score (System Diversity)	0.0020	0.14	0.31	0.09
Perceived Novelty	4.00	3.87	5.00	0.81

Table 4.3: System Diversity & Perceived Novelty Data Statistics

Test	**p-value**	**rho**
Spearmans rank-order correlation	0.045	-0.42

Table 4.4: System Diversity & Perceived Novelty Correlation

As seen from Table 4.4, the ρ **value** is **-0.42** and it is significant ($\mathbf{p = 0.04}, \mathbf{p < 0.05}$). This rejects the null hypothesis and our alternative hypothesis (H_2) is accepted. The ρ value shows that there is a strong negative correlation between the *System Diversity* and the *Perceived Novelty* of the recommended items in both the recommendation lists.

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

To validate our Hypothesis H_3, we ran the Wilcoxon Signed Rank Test to compare the difference (before and after) between the two recommendation lists (**RL1** (before) and **RL2** (after)) in terms of novel items.

Perceived Novelty	Min.	Mean	Max.	Std D.
Before Interaction	1.00	3.13	7.00	1.76
After Interaction	1.00	3.87	8.00	1.91

Table 4.5: Perceived Novelty Data Statistics

Test	p-value
Wilcoxon Signed Rank Test	0.0009

Table 4.6: Perceived Novelty (Pairwise comparison before & after interaction)

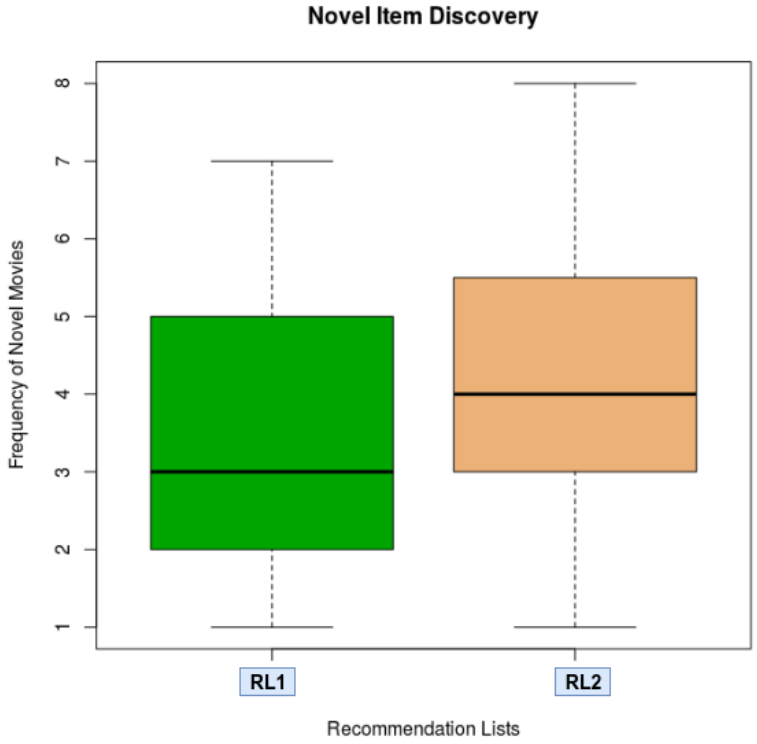


Figure 4.9: Box plot representation Novel Items: Before (RL1) & After (RL2)

Table 4.5 gives more information about the data statistics (min, max, mean, standard deviation values).

As seen from Table 4.6, $\mathbf{p = 0.0009}$ ($\mathbf{p < 0.05}$) and it is significant. This rejects the null hypothesis and our alternative hypothesis ($\mathbf{H_3}$) is accepted. We observe that there is statistical significance in the *Perceived Novelty* of the participants after interacting with the relevant Twitter information. Figure 4.9 shows the frequency of novel movies for both the lists (**RL1** & **RL2**). Analyzing all the data statistics and the relevant information, we infer that the participants did indeed find more *novel* items after interaction with the relevant social information.

Research Question 2 (RQ2)

Does **discovering novel and diverse items** help achieve **better levels of user satisfaction**?

In order to answer **RQ2**, we analyze the impact of *Perceived Novelty* and *Perceived Diversity* on *User Satisfaction*.

We have the following hypotheses for answering **RQ2**:

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

To validate our Hypothesis **H_4**, we ran the Spearman's Rank-Order Correlation test to study the impact of *Perceived Novelty* of the participants on their overall *User Satisfaction*. We compared the overall *perceived novelty* and overall *user satisfaction* scores.

Table 4.7 gives more information about the data statistics (min, max, mean, standard deviation values).

Data Type	**Min.**	**Mean**	**Max.**	**Std D.**
Perceived Novelty	2.67	4.00	4.67	0.53
User Satisfaction	2.00	4.01	5.00	0.60

Table 4.7: Perceived Novelty & User Satisfaction Data Statistics

Test	**p-value**	**rho**
Spearmans rank-order correlation	0.00016	0.70

Table 4.8: Perceived Novelty & User Satisfaction Correlation

As seen from Table 4.8, the ρ **value** is **0.70** and it is significant ($\mathbf{p = 0.00016}$, $\mathbf{p < 0.05}$). This rejects the null hypothesis and our alternative hypothesis (**H_4**) is accepted. The ρ value shows that there is a strong positive correlation between the *Perceived Novelty* and *overall User Satisfaction* of the participants.

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

To validate our Hypothesis **H_5**, we ran the Spearman's Rank-Order Correlation test to study the impact of *Perceived Diversity* of the participants on their overall *User Satisfaction*. We compared the overall *perceived diversity* and *user satisfaction* scores.

Table 4.9 gives more information about the data statistics (min, max, mean, standard deviation values).

Data Type	Min.	Mean	Max.	Std D.
Perceived Novelty	2.25	3.70	4.50	0.56
User Satisfaction	2.00	4.01	4.33	0.60

Table 4.9: Perceived Diversity & User Satisfaction Data Statistics

Test	p-value	rho
Spearmans rank-order correlation	0.003	0.58

Table 4.10: Perceived Diversity & User Satisfaction Correlation

As seen from Table 4.10, the ρ **value** is **0.58** and it is significant ($\mathbf{p = 0.003}$, $\mathbf{p < 0.05}$). This rejects the null hypothesis and our alternative hypothesis ($\mathbf{H_5}$) is accepted. The ρ value shows that there is a strong positive correlation between the *Perceived Diversity* and *overall User Satisfaction* of the participants.

4.4.4 Discussion

In Chapter 3, we described our *MovieTweeters* system in more depth. We discussed all the different phases in detail and defined our dependent variable, *system diversity*, and our independent variables namely *perceived diversity* and *perceived novelty*. Our initial results suggested we were successful in studying and establishing a relationship between the perceived quality attributes (*perceived novelty* & *perceived diversity*) of recommendations and the *satisfaction levels* of the users. We also observed that incorporating relevant social information with an interactive interface did indeed help users discover *novel* items.

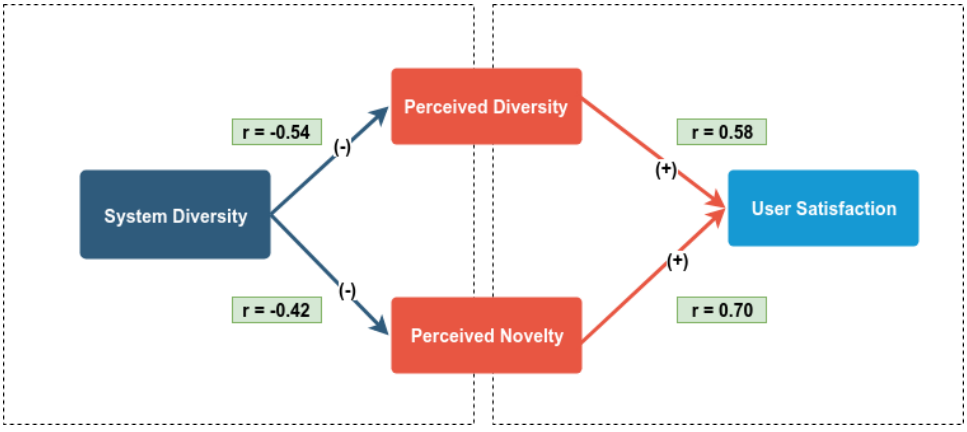


Figure 4.10: System Model

Figure 4.10 shows the relationship between the dependent and the independent variables defined in our study. Based on our results, we see that *System Diversity* negatively influences both *Perceived Diversity* ($\rho = -0.54$) & *Perceived Novelty* ($\rho = -0.42$). We also found out that, *User Satisfaction* was positively influenced by both *Perceived Diversity* ($\rho = 0.58$) & *Perceived Novelty* ($\rho = 0.70$).

While analyzing the impact of *System Diversity* on *Perceived Diversity* & *Perceived Novelty*, we found some surprising results. Notably, the negative correlation between *System Diversity* and *Perceived Diversity*. Following are our observations about what might have contributed to this phenomenon:

- We observed an increase in the *system diversity* scores across the two lists. As the *system diversity* increased, the *perceived diversity* of the users decreased.

    1. **MMR Formula:** We adjusted our MMR formula to balance out *system diversity* and *relevance*. Hence our revised recommendation list was balanced out in terms of *relevance* and *diversity*. This could impact the perception of the participants in a way where their attention was focused more on checking the *relevance* aspect of the recommendations.
    2. **Impact of Popularity:** Another explanation could be the impact of *popularity* on the diversity perception of the users. In post-hoc analysis, we calculated the *Popularity* for a recommendation list taking the average of the top 3 IMDb movie ratings for that list. Appendix B (Table B.6) shows the scores reported for both the recommendation lists for all the participants. We observed that the *popularity* scores actually decreased across the lists. We compared this change to the *perceived diversity* of the participants and found that as the *popularity* decreased across the lists, the *perceived diversity* decreased as well. We state that *popularity* may have played a role in affecting the *perceived diversity* of the participants. Diversification (increase in the *system diversity* across the two lists (**RL1** & **RL2**)) led to recommendation of less popular movies in the second list. It could have been the case where the users were able to recognize the more popular movies in the first list and perceive them to be more diverse than in the second list. Gradually, the *perceived diversity* decreased across the two lists.

- As the *system diversity* increased, the *perceived novelty* of the users decreased.

    1. As our post hoc analysis suggests, increase in the *system diversity* across the two lists lead to decrease in the *popularity* of the recommended items in the second list. We believe this could have had an impact on the *perceived satisfaction* of the users. It could be the case that users easily recognized the more popular movies in the first list and were satisfied with the recommendations also impacting their *perceived novelty*

of the items. On the other hand, less popular movie recommendations in the second list were less recognizable to the users suggesting the decline in their *perceived novelty* levels. However, more research would be needed to study and understand this impact of *perceived satisfaction* and *popularity* on the *perceived novelty* of the users.

Chapter 5

Discussion

In this section, we discuss more in detail about our work with possible ways to improve our approaches. We identify the different limitations in our methods designed to answer our main research objectives.

In order to address issues concerning evaluation of the quality of recommendations and measuring the overall satisfaction of users towards the recommendations, we studied different recommendation system interfaces and the role of social information in recommendation systems. We analyzed the different opportunities, challenges and state of the art in this field. We identified and based our study on two main research directions: First, beyond accuracy metrics (namely *perceived diversity* & *perceived novelty*) for the evaluation of the quality of recommendations. Second, the role of interactive interfaces along with social information as additional input, to improve user satisfaction. In Chapter 1, we identified the following problems in hand:

- *Impact of Interactive Interfaces:* Limited research into how interactive interfaces with social recommendation systems impact the system & perceived quality of recommendations.

- *Study of user satisfaction:* Not enough research done to understand and correlate the impact of user's perceived quality of recommendations on their overall satisfaction.

We defined two main research objectives in order to answer the problems defined above.

- *Impact of System Diversity on Discovery Perception:* Study the influence of an interactive interface (with relevant social information) on user's perception of their recommendations in terms of *novelty* and *diversity*.

- *Impact of Discovery Perception on User Satisfaction:* Study how does an user's perception of their recommendations impact their overall satisfaction.

5.1 Impact of System Diversity on Discovery Perception

Our first research objective was to study the impact of social information (with an interactive interface) on different perceived qualities of recommendations namely *perceived novelty* and *perceived diversity*. We designed an approach where we generated two sets of recommendation lists: First, we generated an initial list of movie recommendations based on the movie preferences learnt from the user. Second, we generated a revised list of movie recommendations based on the relevant social information selected by the user. We made the following two comparisons:

- We first calculated the change in *system diversity* between the two lists. Then we compared this change with the user's *perceived novelty* and *perceived diversity* values of the two recommendation lists.

- We compared the *novel* item discovery for the users before and after interaction with their relevant social information.

For our first comparison, we observed that *system diversity* increased between the two recommendation lists. From the obtained results, we conclude that our interactive interface was successfully able to increase the *system diversity* of the recommendations using the relevant social data. Factors of *inspectability* and *control* played a crucial role as the users were able to explore their social information in a convenient and easy fashion. In addition to previous work done in the field, we state that such an interactive interface can indeed help users explore their neighborhood of relevant items in various domains such as movies, music, professional connections (LinkedIn) and so on. Aligning with the advantages of social recommendations (as stated in Section 2.3.4), using relevant social information helped us increase the diversity and coverage of the recommendations. Based on the selection of relevant social information, different user's opinions and tastes could also be devised.

We also observed that this increase in the *system diversity* had a significant impact on the perceived quality attributes. Prior research in [87] studied the impact of interfaces on *perceived diversity*. They reported increased levels of *perceived diversity* for the users. However in our experiment, even though there was an increase in the *system diversity*, the *perceived diversity* declined for the users. In the context of building an interactive interface for news recommendations, this could be helpful as users would want to get more specific news items to read.

For our second comparison, we observed a positive impact of our interface in terms of helping users discover *novel* content. We observed that in general, users came across more *novel* movies after interacting with their relevant social information. Based on this result, we believe an interactive interface like *MovieTweeters* could also prove useful to help users come out of their filter-bubble and explore more new and novel content.

The type of social information also plays an important role while designing such an interface. For our study, our social data facilitated us in making better recommendations. It could be interesting to see how such an interface with any other specific type of social data can help study user relations such as *trust relations*.

5.2 Impact of Discovery Perception on User Satisfaction

For our second research objective, our approach analyzed the impact of user discovery perception on user satisfaction. From our obtained results, we state that users who perceived the discovery of more *novel* and *diverse* movies via the interface, were more satisfied with the recommendations and also with their overall experience of the interface. This could act as a precedent for developing social recommendation interfaces with user satisfaction and content discovery being the primary goals.

Table 5.1 shows an overview of the results obtained. We list all the obtained significant results with their correlation values (ρ). As seen from the table, perceived qualities such as *perceived novelty* & *perceived diversity* had a considerable positive impact on the *user satisfaction* levels.

Research Question	**Hypothesis**	**Variable Compared**	**Significant**	**ρ value**
RQ1	H1	System Diversity - Perceived Diversity	✓	-0.54
RQ1	H2	System Diversity - Perceived Novelty	✓	-0.42
RQ1	H3	Perceived Novelty	✓	-
RQ2	H4	Perceived Novelty - User Satisfaction	✓	0.70
RQ2	H5	Perceived Diversity - User Satisfaction	✓	0.58

Table 5.1: Summarization of the obtained results

5.3 Limitations

In this section, we discuss the limitations in our system. We identify the following main limitations in our study:

- **Limited recommendation algorithmic approach:** As the main focus of this thesis was on the interface aspect of recommendation systems, we used only one algorithm (item-item based collaborative filtering) to generate the initial set of movie recommendations.

- **MMR λ value:** For our study, we defined an approach where we considered only one value of λ in our MMR formula (Refer Section 3.5.2). We did

this because we wanted to achieve a balance between the *relevance* and the *diversity* for the items in the revised recommendation list.

- **Limited feature selection:** Throughout our study, we used only the genre features of the movies while comparing Twitter Users to system users. The genre features provided us a good starting point to compare the consumption of the system users. But it also limited the scope of the study as we developed our interface only based on the genre feature. To adapt our approach to build similar systems in other domains, research would need to be done in order to select the right features.

- **Limited sample size:** In our study, we had only a limited number of participants (n=23). While this group size was more accessible, it might have hindered our ability to generalize our results to a much bigger population (consisting of different groups of users and ages).

- **Exclusion of personality traits:** The scope of our study was mainly focused on the impact of an interactive interface (with relevant social information) on the quality of recommendations (in terms of *novelty* and *diversity*). While we measured the user's *perceived diversity* of the recommended items, we did not take into consideration the individualistic traits of the user such as their personality. This phenomenon could potentially also have been an influencing factor for the surprising way our *system diversity* influenced the *perceived diversity* of the users.

Chapter 6

Conclusion

In this thesis, we research about the impact of interactive recommendation system interfaces on the quality of recommendations. We also understand and analyze the impact of discovery perception on user satisfaction. To answer our defined research questions (Section 1.2), we started by first understanding the background of social recommendation systems and the different beyond accuracy evaluation metrics. An extensive literature study provided us with an insight into different recommendation systems interface aspects such as *Inspectability & Control* and their impact on users.

Primarily, the previous research in recommendation systems has mainly focused on evaluating the recommendations based on their accuracy. More recent research has begun evaluating the subjective perception of the recommendations and has developed additional attributes such as *diversity* and *novelty* to evaluate recommendation systems. Also, not enough research had been done in the field to understand and correlate the impact of perceived attributes of recommendations such as *perceived diversity* and *perceived novelty* on overall *user satisfaction*. Based on the identified research gaps, we formulated our two main research objectives: First, to study the impact of an interactive interface with social information on the quality of recommendations and discovery perception of the users. Second, the understand and correlate the impact of discovery perception on overall user satisfaction.

To answer our research objectives, we introduced and evaluated a novel interactive interface, *MovieTweeters*. It is a movie recommender system which combines social information with a traditional recommendation algorithm. This allowed us to generate recommendations that were both current (since the social information is constantly updating), and novel. We conducted offline and online evaluations to test our interface.

We conclude that our interactive interface with social information did influence and impact the way users perceived their recommendations. Our interface proved that incorporation of relevant social information with an interactive interface does

indeed help users discover more *novel* items. It also provided valuable insight into the relationship between how users perceive the recommended items and their overall satisfaction. We observed that users who *perceived* that they discovered more *novel* and *diverse* items also reported increased levels of *user satisfaction*. Even though we successfully were able to increase the *system diversity* of the recommendations, it had a negative correlation with users perception of *novelty* and *diversity* of the items.

6.1 Future Work

Our experiments show that an interactive interface like *MovieTweeters* is successfully able to help users discover more *novel* items and also provided insight into how increased levels of *perceived diversity* and *perceived novelty* lead to increased levels of *user satisfaction*. Future works could potentially overcome some of the limitations of our system and eventually extend our system to generate better recommendations irrespective of the domain (music, news and so on). In this section, we discuss about the extension of this research in certain aspects:

- **Comparing different algorithmic approaches:** An experiment comparing different recommendation algorithms for recommendations could prove helpful to understand the impact these different recommendation algorithms have on perceived quality of the recommendations and on overall user satisfaction.

- **Bigger sample size:** Larger user experiment with different groups of users and ages could be performed. It would be interesting to study how such a sample size affects the way users perceive the quality of recommendations and their overall user satisfaction.

- **Inclusion of more features:** The inclusion of other movies specific features (e.g., features such as box office revenues, year/era of release) could provide more insight into how two users are correlated which would also impact the final recommendation list.

  In the case of music or news, further research would be required to understand which features are the more useful in understanding user behavior better and eventually generating recommendations which make them more satisfied.

- **MMR (λ) values:** Multiple experiments with different MMR (λ) values could help us understand its impact on the revised recommendation list and ultimately on the perception of the users and their overall user satisfaction.

Bibliography

[1] S. M. McNee, J. Riedl and J. 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*, ACM, 2006, pp. 1097–1101.

[2] J. L. Herlocker, J. A. Konstan, A. Borchers and J. Riedl, 'An algorithmic framework for performing collaborative filtering', in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 1999, pp. 230–237.

[3] T. Tran and R. Cohen, 'Hybrid recommender systems for electronic commerce', in *Proc. Knowledge-Based Electronic Markets, Papers from the AAAI Workshop, Technical Report WS-00-04, AAAI Press*, 2000.

[4] S. Bostandjiev, J. O'Donovan and T. Höllerer, 'Tasteweights: A visual interactive hybrid recommender system', in *Proceedings of the sixth ACM conference on Recommender systems*, ACM, 2012, pp. 35–42.

[5] X. Su and T. M. Khoshgoftaar, 'A survey of collaborative filtering techniques', *Advances in artificial intelligence*, vol. 2009, p. 4, 2009.

[6] P. V. Marsden and N. E. Friedkin, 'Network studies of social influence', *Sociological Methods & Research*, vol. 22, no. 1, pp. 127–151, 1993.

[7] I. King, M. R. Lyu and H. Ma, 'Introduction to social recommendation', in *Proceedings of the 19th international conference on World wide web*, ACM, 2010, pp. 1355–1356.

[8] H. Kautz, B. Selman and M. Shah, 'Referral web: Combining social networks and collaborative filtering', *Communications of the ACM*, vol. 40, no. 3, pp. 63–65, 1997.

[9] J. Tang, X. Hu and H. Liu, 'Social recommendation: A review', *Social Network Analysis and Mining*, vol. 3, no. 4, pp. 1113–1133, 2013.

[10] J. Golbeck, 'Generating predictive movie recommendations from trust in social networks', *Trust Management*, pp. 93–104, 2006.

[11] P. Massa and P. Avesani, 'Trust-aware recommender systems', in *Proceedings of the 2007 ACM conference on Recommender systems*, ACM, 2007, pp. 17–24.

[12] N. Ma, E.-P. Lim, V.-A. Nguyen, A. Sun and H. Liu, 'Trust relationship prediction using online product review data', in *Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management*, ACM, 2009, pp. 47–54.

[13] M. Jamali and M. Ester, 'Trustwalker: A random walk model for combining trust-based and item-based recommendation', in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2009, pp. 397–406.

[14] H. Ma, D. Zhou, C. Liu, M. R. Lyu and I. King, 'Recommender systems with social regularization', in *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM, 2011, pp. 287–296.

[15] I. Guy, 'Social recommender systems', in *Recommender Systems Handbook*, Springer, 2015, pp. 511–543.

[16] W.-Y. Chen, J.-C. Chu, J. Luan, H. Bai, Y. Wang and E. Y. Chang, 'Collaborative filtering for orkut communities: Discovery of user latent behavior', in *Proceedings of the 18th international conference on World wide web*, ACM, 2009, pp. 681–690.

[17] B. Sigurbjörnsson and R. Van Zwol, 'Flickr tag recommendation based on collective knowledge', in *Proceedings of the 17th international conference on World Wide Web*, ACM, 2008, pp. 327–336.

[18] H. Ma, T. C. Zhou, M. R. Lyu and I. King, 'Improving recommender systems by incorporating social contextual information', *ACM Transactions on Information Systems (TOIS)*, vol. 29, no. 2, p. 9, 2011.

[19] M. Jiang, P. Cui, R. Liu, Q. Yang, F. Wang, W. Zhu and S. Yang, 'Social contextual recommendation', in *Proceedings of the 21st ACM international conference on Information and knowledge management*, ACM, 2012, pp. 45–54.

[20] C. Kofler, L. Caballero, M. Menendez, V. Occhialini and M. Larson, 'Near2me: An authentic and personalized social media-based recommender for travel destinations', in *Proceedings of the 3rd ACM SIGMM international workshop on Social media*, ACM, 2011, pp. 47–52.

[21] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl, 'Grouplens: An open architecture for collaborative filtering of netnews', in *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, ACM, 1994, pp. 175–186.

[22] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, 'Item-based collaborative filtering recommendation algorithms', in *Proceedings of the 10th international conference on World Wide Web*, ACM, 2001, pp. 285–295.

[23] P. Massa, P. Avesani *et al.*, 'Trust-aware collaborative filtering for recommender systems', *CoopIS/DOA/ODBASE (1)*, vol. 3290, pp. 492–508, 2004.

[24] P. Massa and P. Avesani, 'Controversial users demand local trust metrics: An experimental study on epinions. com community', in *AAAI*, vol. 5, 2005, pp. 121–126.

[25] C. Basu, H. Hirsh, W. Cohen *et al.*, 'Recommendation as classification: Using social and content-based information in recommendation', in *Aaai/iaai*, 1998, pp. 714–720.

[26] H. Ma, H. Yang, M. R. Lyu and I. King, 'Sorec: Social recommendation using probabilistic matrix factorization', in *Proceedings of the 17th ACM conference on Information and knowledge management*, ACM, 2008, pp. 931–940.

[27] L. Hong, A. S. Doumith and B. D. Davison, 'Co-factorization machines: Modeling user interests and predicting individual decisions in twitter', in *Proceedings of the sixth ACM international conference on Web search and data mining*, ACM, 2013, pp. 557–566.

[28] M. Jamali and M. Ester, 'A matrix factorization technique with trust propagation for recommendation in social networks', in *Proceedings of the fourth ACM conference on Recommender systems*, ACM, 2010, pp. 135–142.

[29] Q. Yuan, S. Zhao, L. Chen, Y. Liu, S. Ding, X. Zhang and W. Zheng, 'Augmenting collaborative recommender by fusing explicit social relationships', in *Workshop on Recommender Systems and the Social Web, Recsys 2009*, 2009.

[30] P. Symeonidis, E. Tiakas and Y. Manolopoulos, 'Product recommendation and rating prediction based on multi-modal social networks', in *Proceedings of the fifth ACM conference on Recommender systems*, ACM, 2011, pp. 61–68.

[31] J. L. Herlocker, J. A. Konstan and J. Riedl, 'Explaining collaborative filtering recommendations', in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, ACM, 2000, pp. 241–250.

[32] N. Tintarev and J. Masthoff, 'Evaluating the effectiveness of explanations for recommender systems', *User Modeling and User-Adapted Interaction*, vol. 22, no. 4, pp. 399–439, 2012.

[33] ——, 'Designing and evaluating explanations for recommender systems', *Recommender Systems Handbook*, pp. 479–510, 2011.

[34] B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan and A. Kobsa, 'Inspectability and control in social recommenders', in *Proceedings of the sixth ACM conference on Recommender systems*, ACM, 2012, pp. 43–50.

[35] N. Tintarev and J. Masthoff, 'Explaining recommendations: Design and evaluation', in *Recommender Systems Handbook*, Springer, 2015, pp. 353–382.

[36] R. R. Sinha, K. Swearingen *et al.*, 'Comparing recommendations made by online systems and friends.', in *DELOS workshop: personalisation and recommender systems in digital libraries*, vol. 106, 2001.

[37] B. Kang, N. Tintarev, T. Höllerer and J. OfffdfffdfffdDonovan, 'What am i not seeing? an interactive approach to social content discovery in microblogs', in *International Conference on Social Informatics*, Springer, 2016, pp. 279–294.

[38] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden and R. C. Miller, 'Twitinfo: Aggregating and visualizing microblogs for event exploration', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 2011, pp. 227–236.

[39] J. Schaffer, P. Giridhar, D. Jones, T. Höllerer, T. Abdelzaher and J. O'Donovan, 'Getting the message?: A study of explanation interfaces for microblog data analysis', in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, ACM, 2015, pp. 345–356.

[40] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam and E. H. Chi, 'Eddi: Interactive topic-based browsing of social status streams', in *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, ACM, 2010, pp. 303–312.

[41] B. Gretarsson, J. O'Donovan, S. Bostandjiev, C. Hall and T. Höllerer, 'Smallworlds: Visualizing social recommendations', in *Computer Graphics Forum*, Wiley Online Library, vol. 29, 2010, pp. 833–842.

[42] I. Guy, I. Ronen and E. Wilcox, 'Do you know?: Recommending people to invite into your social network', in *Proceedings of the 14th international conference on Intelligent user interfaces*, ACM, 2009, pp. 77–86.

[43] S. Bostandjiev, J. O'Donovan and T. Höllerer, 'Linkedvis: Exploring social and semantic career recommendations', in *Proceedings of the 2013 international conference on Intelligent user interfaces*, ACM, 2013, pp. 107–116.

[44] K. Verbert, D. Parra, P. Brusilovsky and E. Duval, 'Visualizing recommendations to support exploration, transparency and controllability', in *Proceedings of the 2013 international conference on Intelligent user interfaces*, ACM, 2013, pp. 351–362.

[45] B. Lamche, U. Adıgüzel and W. Wörndl, 'Interactive explanations in mobile shopping recommender systems', in *Joint Workshop on Interfaces and Human Decision Making in Recommender Systems*, 2014, p. 14.

[46] L. Chen and P. Pu, 'Critiquing-based recommenders: Survey and emerging trends', *User Modeling and User-Adapted Interaction*, vol. 22, no. 1, pp. 125–150, 2012.

[47] K. McCarthy, Y. Salem and B. Smyth, 'Experience-based critiquing: Reusing critiquing experiences to improve conversational recommendation', in *International Conference on Case-Based Reasoning*, Springer, 2010, pp. 480–494.

[48] P. Pu, L. Chen and P. Kumar, 'Evaluating product search and recommender systems for e-commerce environments', *Electronic Commerce Research*, vol. 8, no. 1-2, pp. 1–27, 2008.

[49] J. O'Donovan, B. Gretarsson, S. Bostandjiev, T. Hollerer and B. Smyth, 'A visual interface for social information filtering', in *Computational Science and Engineering, 2009. CSE'09. International Conference on*, IEEE, vol. 4, 2009, pp. 74–81.

[50] J. O'Donovan, B. Smyth, B. Gretarsson, S. Bostandjiev and T. Höllerer, 'Peerchooser: Visual interactive recommendation', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2008, pp. 1085–1088.

[51] S. McNee, S. Lam, J. Konstan and J. Riedl, 'Interfaces for eliciting new user preferences in recommender systems', *User Modeling 2003*, pp. 148–148, 2003.

[52] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu and C. Newell, 'Explaining the user experience of recommender systems', *User Modeling and User-Adapted Interaction*, vol. 22, no. 4-5, pp. 441–504, 2012.

[53] C. Gena, R. Brogi, F. Cena and F. Vernero, 'The impact of rating scales on userffffdffffdfffds rating behavior', in *International Conference on User Modeling, Adaptation, and Personalization*, Springer, 2011, pp. 123–134.

[54] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan and J. Riedl, 'Is seeing believing?: How recommender system interfaces affect users' opinions', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 2003, pp. 585–592.

[55] K. Swearingen and R. Sinha, 'Interaction design for recommender systems', in *Designing Interactive Systems*, vol. 6, 2002, pp. 312–334.

[56] P. Pu, L. Chen and R. Hu, 'A user-centric evaluation framework for recommender systems', in *Proceedings of the fifth ACM conference on Recommender systems*, ACM, 2011, pp. 157–164.

[57] S. Castagnos, N. Jones and P. Pu, 'Recommenders' influence on buyers' decision process', in *Proceedings of the third ACM conference on Recommender systems*, ACM, 2009, pp. 361–364.

[58] N. Jones, P. Pu and L. Chen, 'How users perceive and appraise personalized recommendations', *User Modeling, Adaptation, and Personalization*, pp. 461–466, 2009.

[59] J. Schaffer, T. Höllerer and J. O'Donovan, 'Hypothetical recommendation: A study of interactive profile manipulation behavior for recommender systems.', in *FLAIRS Conference*, 2015, pp. 507–512.

[60] P. Pu and L. Chen, 'Trust-inspiring explanation interfaces for recommender systems', *Knowledge-Based Systems*, vol. 20, no. 6, pp. 542–556, 2007.

[61] ——, 'Trust building with explanation interfaces', in *Proceedings of the 11th international conference on Intelligent user interfaces*, ACM, 2006, pp. 93–100.

[62] G. Shani and A. Gunawardana, 'Evaluating recommendation systems', *Recommender systems handbook*, pp. 257–297, 2011.

[63] J. L. Herlocker, J. A. Konstan, L. G. Terveen and J. T. Riedl, 'Evaluating collaborative filtering recommender systems', *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 5–53, 2004.

[64] K. Swearingen and R. Sinha, 'Beyond algorithms: An hci perspective on recommender systems', in *ACM SIGIR 2001 Workshop on Recommender Systems*, vol. 13, 2001, pp. 1–11.

[65] L. Zhang, 'The definition of novelty in recommendation system.', *Journal of Engineering Science & Technology Review*, vol. 6, no. 3, 2013.

[66] Ò. Celma Herrada *et al.*, *Music recommendation and discovery in the long tail*. Universitat Pompeu Fabra, 2009.

[67] S. Vargas and P. Castells, 'Rank and relevance in novelty and diversity metrics for recommender systems', in *Proceedings of the fifth ACM conference on Recommender systems*, ACM, 2011, pp. 109–116.

[68] P. Castells, S. Vargas and J. Wang, 'Novelty and diversity metrics for recommender systems: Choice, discovery and relevance', 2011.

[69] Ò. Celma and P. Herrera, 'A new approach to evaluating novel recommendations', in *Proceedings of the 2008 ACM conference on Recommender systems*, ACM, 2008, pp. 179–186.

[70] C.-N. Ziegler, S. M. McNee, J. A. Konstan and G. Lausen, 'Improving recommendation lists through topic diversification', in *Proceedings of the 14th international conference on World Wide Web*, ACM, 2005, pp. 22–32.

[71] B. Smyth and P. McClave, 'Similarity vs. diversity', *Case-Based Reasoning Research and Development*, pp. 347–361, 2001.

[72] M. Ge, F. Gedikli and D. Jannach, 'Placing high-diversity items in top-n recommendation lists', in *Workshop chairs*, 2011, p. 65.

[73] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling and Y.-C. Zhang, 'Solving the apparent diversity-accuracy dilemma of recommender systems', *Proceedings of the National Academy of Sciences*, vol. 107, no. 10, pp. 4511–4515, 2010.

[74] N. Lathia, S. Hailes, L. Capra and X. Amatriain, 'Temporal diversity in recommender systems', in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2010, pp. 210–217.

[75] B. Ferwerda, M. P. Graus, A. Vall, M. Tkalcic and M. Schedl, 'How item discovery enabled by diversity leads to increased recommendation list attractiveness', in *Proceedings of the Symposium on Applied Computing*, ACM, 2017, pp. 1693–1696.

[76] Y. Zhu, Y. Lan, J. Guo and X. Cheng, 'Structural learning of diverse ranking', *arXiv preprint arXiv:1504.04596*, 2015.

[77] M. Zhang and N. Hurley, 'Avoiding monotony: Improving the diversity of recommendation lists', in *Proceedings of the 2008 ACM conference on Recommender systems*, ACM, 2008, pp. 123–130.

[78] G. Adomavicius and Y. Kwon, 'Toward more diverse recommendations: Item re-ranking methods for recommender systems', in *Workshop on Information Technologies and Systems*, Citeseer, 2009.

[79] J. Carbonell and J. Goldstein, 'The use of mmr, diversity-based reranking for reordering documents and producing summaries', in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 1998, pp. 335–336.

[80] S. Xie and Y. Liu, 'Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization', in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, IEEE, 2008, pp. 4985–4988.

[81] F. M. Harper and J. A. Konstan, 'The movielens datasets: History and context', *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 5, no. 4, p. 19, 2016.

[82] A. M. Rashid, I. Albert, D. Cosley, S. K. Lam, S. M. McNee, J. A. Konstan and J. Riedl, 'Getting to know you: Learning new user preferences in recommender systems', in *Proceedings of the 7th international conference on Intelligent user interfaces*, ACM, 2002, pp. 127–134.

[83] Y.-J. Park and A. Tuzhilin, 'The long tail of recommender systems and how to leverage it', in *Proceedings of the 2008 ACM conference on Recommender systems*, ACM, 2008, pp. 11–18.

[84] S. Dooms, T. De Pessemier and L. Martens, 'Movietweetings: A movie rating dataset collected from twitter', in *Workshop on Crowdsourcing and human computation for recommender systems, CrowdRec at RecSys*, vol. 2013, 2013, p. 43.

[85] J. F. Forst, A. Tombros and T. Roelleke, 'Less is more: Maximal marginal relevance as a summarisation feature', in *Conference on the Theory of Information Retrieval*, Springer, 2009, pp. 350–353.

[86] K. Bradley and B. Smyth, 'Improving recommendation diversity', in *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland*, Citeseer, 2001, pp. 85–94.

[87] R. Hu and P. Pu, 'Enhancing recommendation diversity with organization interfaces', in *Proceedings of the 16th international conference on Intelligent user interfaces*, ACM, 2011, pp. 347–350.

Appendix A

User Demographics Questionnaire

The following questions were asked to the participants to get their relevant user demographic information.

1. Gender
    - Male
    - Female
    - Prefer not to answer
2. Age
    - 18-24
    - 25-34
    - 35-44
    - 45-49
    - 50-55
    - 56+
3. We would like to understand your movie consumption behavior. How many movies do you watch every month?
    - I rarely watch movies.
    - 1-3
    - 4-7
    - More than 10.
4. Which of the following options below describe you the best? (You can select more than one option)

- **Casual:** I do not watch many movies per year and watching movies is not my hobby. I usually prefer to spend my time doing something else rather than watching movies.
- **Average Movie Watcher:** I watch several movies each year including some Oscar contenders, some seasonal classics (Christmas movies etc.). I don't really consider watching movies as my hobby. I consider it just as a fun past time.
- **Genre Specific:** I am mainly into movies from specific genres. I watch a lot of movies from my favorite genres and like to explore different sub-genres within my favorite genres (for example, Star Wars and Black Mirror both explore and belong to different themes).
- **Movie Buff:** Watching movies is my hobby and my taste is relatively varied. I like to watch the old "classics" and usually check out the best and promising movies each year. I like film as an art form and love exploring new and different flavors of films.
- **None of the above?** Select this option and feel free to use the space below to describe yourself.

Appendix B

Result Data

B.1 Perceived Novelty

The data collected during the online experiment for *Perceived Novelty*. The overall score was calculated as an average of all the response to Q1, Q2 and Q3. All responses to the questions (Q1 to Q3) were in the form of a Likert scale (1 - 5).

Participant ID	Novel Before (RL1)	Novel After (RL2)	Q 1	Q 2	Q 3	Overall Score
70000	7	8	4	4	3	3.67
70001	1	2	5	5	4	4.67
70002	1	2	5	5	4	4.67
70003	2	4	4	4	4	4.00
70004	3	6	5	4	5	4.67
70005	4	5	4	4	4	4.00
70006	3	3	4	5	4	4.33
70007	2	1	4	4	3	3.67
70008	5	6	5	4	4	4.33
70009	2	3	4	4	4	4.00
70010	4	6	3	5	5	4.33
70011	2	4	4	5	5	4.67
70013	4	5	2	5	4	3.67
70014	5	7	2	5	4	3.67
70015	1	3	2	4	4	3.33
70016	5	5	4	4	4	4.00
70017	1	5	4	5	4	4.33
70018	2	3	4	4	5	4.33
70019	5	5	5	4	4	4.33
70020	1	3	5	5	2	4.00
70021	2	1	4	2	3	3.00
70022	5	4	2	4	1	2.33
70023	5	7	3	4	4	3.67

Table B.1: Perceived Novelty Scores

B.2 Perceived Diversity

The data collected during the online experiment for *Perceived Diversity*:

Participant ID	Q 1	Q 2	Q 3	Q 4	Overall Score
70000	4	3	4	4	3.75
70001	5	4	5	4	4.50
70002	3	1	3	4	2.75
70003	4	3	4	5	4.00
70004	5	2	5	5	4.25
70005	5	4	4	4	4.25
70006	4	2	4	4	3.50
70007	4	2	3	4	3.25
70008	4	4	4	4	4.00
70009	4	3	4	3	3.50
70010	5	4	4	5	4.50
70011	4	3	4	4	3.75
70013	4	4	4	4	4.00
70014	4	4	4	4	4.00
70015	4	5	4	4	4.25
70016	4	2	4	4	3.50
70017	5	2	4	4	3.75
70018	5	3	4	4	4.00
70019	4	2	4	3	3.25
70020	4	2	4	4	3.50
70021	3	1	3	2	2.25
70022	5	1	2	3	2.75
70023	4	4	4	5	4.25

Table B.2: Perceived Diversity Scores

The overall score was calculated as an average of all the response to Q1, Q2, Q3 and Q4. All responses to the questions (Q1 to Q4) were in the form of a Likert scale (1 - 5).

B.3 User Satisfaction

The overall score was calculated as an average of all the response to Q1, Q2, Q3, Q4, Q5 and Q6. All responses were in the form of a Likert scale (1 - 5).

Participant ID	Q 1	Q 2	Q 3	Q 4	Q 5	Q 6	Overall Score
70000	4	4	4	4	4	4	4.00
70001	5	3	5	5	5	4	4.50
70002	4	4	3	4	5	5	4.17
70003	4	4	4	5	4	4	4.17
70004	5	5	5	5	5	5	5.00
70005	4	4	5	4	4	4	4.17
70006	5	4	4	5	5	5	4.67
70007	4	3	4	4	4	4	3.83
70008	5	4	4	4	4	5	4.33
70009	4	4	3	4	4	4	3.83
70010	5	4	4	5	4	5	4.50
70011	4	3	4	5	5	4	4.17
70013	4	5	4	4	4	4	4.17
70014	4	4	4	3	3	5	3.83
70015	3	2	4	4	5	4	3.67
70016	4	3	4	4	3	4	3.67
70017	4	4	4	4	4	5	4.17
70018	5	4	4	5	4	5	4.50
70019	4	5	3	4	4	4	4.00
70020	4	4	2	4	4	4	3.67
70021	2	2	3	4	3	4	3.00
70022	1	4	4	1	1	1	2.00
70023	4	4	5	4	5	4	4.33

Table B.3: User Satisfaction Scores

B.4 System Novelty

System Novelty scores achieved for each participant were as follows:

Participant ID	System Novelty
70000	1
70001	0.9
70002	0.7
70003	0.8
70004	0.8
70005	0.9
70006	0.8
70007	0.9
70008	0.9
70009	0.9
70010	0.9
70011	0.9
70013	0.9
70014	0.9
70015	1
70016	0.8
70017	0.9
70018	0.9
70019	0.9
70020	0.6
70021	0.7
70022	1
70023	0.9

Table B.4: System Novelty Scores

B.5 System Diversity

System Diversity scores achieved for each participant were as follows:

Participant	RL1	RL2	Δ Score (System Diversity)
70000	0.608	0.736	0.128
70001	0.497	0.785	0.288
70002	0.497	0.709	0.212
70003	0.655	0.747	0.092
70004	0.698	0.587	0.111
70005	0.441	0.637	0.196
70006	0.771	0.677	0.094
70007	0.527	0.381	0.146
70008	0.466	0.659	0.193
70009	0.641	0.765	0.124
70010	0.527	0.549	0.022
70011	0.613	0.679	0.066
70013	0.456	0.741	0.285
70014	0.482	0.740	0.258
70015	0.554	0.524	0.030
70016	0.438	0.615	0.177
70017	0.664	0.614	0.050
70018	0.664	0.662	0.002
70019	0.749	0.712	0.037
70020	0.425	0.691	0.266
70021	0.425	0.737	0.312
70022	0.657	0.618	0.039
70023	0.657	0.762	0.105

Table B.5: System Diversity Scores

B.6 Popularity Scores

The popularity scores calculated for each user for both the lists (RL1 & RL2) are shown in Table B.6. Δ score was calculated by taking the absolute difference between popularity values of the two lists.

Participant ID	Popularity (RL1)	Popularity (RL2)	Δ Popularity Score
70000	8.70	8.00	0.70
70001	8.70	8.26	0.43
70002	8.66	8.70	0.03
70003	8.70	8.60	0.10
70004	8.76	8.20	0.56
70005	8.43	7.90	0.53
70006	8.76	8.66	0.10
70007	8.70	8.13	0.56
70008	8.70	8.23	0.46
70009	8.66	8.36	0.30
70010	8.70	9.00	0.30
70011	8.66	8.53	0.13
70013	8.66	8.83	0.16
70014	8.66	8.40	0.26
70015	8.66	8.56	0.10
70016	8.70	8.53	0.16
70017	8.80	8.03	0.76
70018	8.70	8.53	0.16
70019	8.76	8.33	0.43
70020	8.70	8.90	0.20
70021	8.70	8.66	0.03
70022	8.66	8.33	0.33
70023	8.76	8.60	0.16

Table B.6: Popularity Scores