

A Test Collection for Offline Evaluation of Recommender Systems

by

Houmaan Chamani

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Management Sciences

Waterloo, Ontario, Canada, 2024

© Houmaan Chamani 2024

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Professor Mark D. Smucker had a direct contribution to the ideation of this work, helped write the code used to train and produce recommendations, developed the process for and did the selection of parameter values for some of the recommendation algorithms, wrote programs and assisted with the writing of other programs, and provided assistance and code that aided in part for the analysis of the collected data, and helped write the abstract and edit the thesis. I also received help from Abdul Manaam in early filtering of the Movielens dataset from TV-shows and Series. The remainder of this thesis, which I wrote with Professor Mark D. Smucker's supervision, is entirely my own.

Abstract

Recommendation systems have long been evaluated by collecting a large number of individuals' ratings for items, and then dividing these ratings into test and train sets to see how well recommendation algorithms can predict individuals' preferences. A complaint about this approach is that the evaluation measures can only use a small number of known preferences and have no information about the majority of recommended items. Prior research has shown that offline evaluation of recommendation systems using a test/train split methodology may not agree with actual user preferences when all recommended items are judged by the user. To address this issue, we apply traditional information retrieval test collection construction techniques for movie recommendations. An information retrieval test collection is composed of documents, search topics, and relevance judgments that tell us which documents are relevant for each topic. For our test collection, each search topic is an individual who is looking for movies to watch. In other words, while the search topic is always "Please recommend me movies that I will be interested in watching," the context of the search topic changes to be the individual who is requesting the recommendations. When document collections are too large to be completely judged by assessors, the traditional approach is to use pooling. We followed this same approach in the construction of our test collection. For each individual, we used their existing profile of rated movies as input to a wide range of recommendation algorithms to produce recommendations for movies not found in their profile. We then pooled these recommendations separately for each person and asked them to rate the movies. In addition to rating, we also had each individual rate a random sample of movies selected from their ratings profile to measure their consistency in rating. The resulting new test collection consists of 51 individual ratings profiles totaling 123,104 ratings and 31,236 relevance judgments. In this thesis, we detail the creation of the test collection and provide an analysis of the individuals that comprise its search topics, and we analyze the collection's relevance judgments as well as other aspects.

Acknowledgements

I would like to express my gratitude to my supervisor, Professor Mark D. Smucker, for his support and invaluable insights throughout this research journey.

I extend my sincere appreciation to Professor Charles L. A. Clarke and Professor Olga Vechtomova for their effort in reading this thesis.

I am also grateful to the GroupLens research group at the University of Minnesota for their generous support. Special thanks go to Daniel Kluver and Professor Joseph Konstan for allowing us to display our recruitment banner on the MovieLens website and their creation of the ML-32M dataset. Additionally, I would like to thank Michael D. Ekstrand for his direct support with the LensKit package.

I would also like to acknowledge the developers of the RecBole framework. Their work has greatly contributed to the field of recommender systems and has been a valuable resource for this study.

This research was supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2020-04665, RRGPAS2020-0008). I am thankful for their financial support, which made this work possible.

Dedication

This is dedicated to my family and fabulous friends.

Table of Contents

Author’s Declaration	ii
Statement of Contributions	iii
Abstract	iv
Acknowledgements	v
Dedication	vi
List of Figures	xii
List of Tables	xiv
1 Introduction	1
1.1 Research Motivation	1
1.2 Thesis Overview	6
1.3 Contributions	6

2	Related Work	7
2.1	Test Collections in Information Retrieval	7
2.2	Evaluation in Information Retrieval	8
2.3	Recommender Systems	8
2.3.1	Online Evaluation	9
2.3.2	Offline Evaluation	10
2.4	Advancements in Offline Evaluation Methodology	11
2.5	MovieLens Datasets	12
2.5.1	MovieLens Versions and Use Cases	13
3	Creating the Test Collection	14
3.1	Dataset Preparation	14
3.1.1	Dataset Extension with Movie Details	15
3.1.2	Base Dataset Filtering	15
3.2	Recommendation Algorithms	16
3.2.1	Recommendation Frameworks	17
3.2.2	Algorithm interface	17
3.2.3	Fine-Tuning RecBole Algorithms	19
3.2.4	Fine-Tuning LensKit Algorithms	24
3.3	Participant Recruitment	28
3.4	Website Design and Study Pipeline	29
3.4.1	Screening Process and Data Obtaining	30

3.4.2	Removing Potentially Existing Profiles	30
3.4.3	Dataset Extension with Participant Profiles	31
3.4.4	Explicit and Implicit Datasets	32
3.4.5	Recommendation Generation	33
3.4.6	Tutorial and Knowledge Check	35
3.4.7	Phase one	35
3.4.8	Consistency Check	37
3.4.9	Phase Two	39
3.5	Test Collection Data Cleaning	39
3.5.1	Study Funnel	40
3.5.2	Data Preparation	42
4	Participant Judgment Analysis	43
4.1	Participant Demographics	43
4.1.1	Age Statistics	43
4.1.2	Gender Statistics	44
4.1.3	Ethnicity Statistics	46
4.1.4	Education Statistics	47
4.2	Nature of Movielens Users & Participants	48
4.2.1	Number of Ratings Per User	48
4.2.2	Distribution of Rating Values	50
4.2.3	Distribution of Average Ratings	53

4.2.4	Last Year of Movie Rated	53
4.2.5	Year Distribution of Rated Movies	56
4.2.6	User Lifespan	56
4.2.7	Implications	58
4.3	Consistency Checks	60
4.3.1	Test-Retest Reliability	61
4.3.2	Rate of Past Views for Consistency Items	62
4.4	Participant Feedback Analysis	63
4.5	Pooling Depth & Number of Judgements	64
4.6	Data & Participant Insights	66
4.6.1	Interest for Unseen Movies	66
4.6.2	Interest for Seen, Absent from Movielens	68
4.6.3	Interest for Seen, Consistency Check Movies	69
4.6.4	Interest for All Seen Movies	70
4.6.5	Interest for Seen vs Unseen Movies	71
4.6.6	Interest vs Familiarity	73
4.6.7	Interest with Low Predicted Rating	74
4.6.8	Interest vs Popularity	76
4.6.9	Interest vs Recency of Movies	76
4.6.10	Participant Rate of Judgement	77
4.6.11	Factors Influencing Interest Level	79

5 Conclusion	82
References	84
APPENDICES	92
A Web Interface Used for User Study	93
A.1 Overview and Signup	94
A.2 Login	95
A.3 Homepage	96
A.4 Movie Judging	96
A.4.1 Judging Speed and Use of other Websites	98
A.5 List of Judged Movies	99
A.6 Judgement Completion	99
A.7 Ranking	100
A.8 Ranking Unavailable	101
A.9 Post study questionnaire	102
A.10 Final page	103
B Web Interface Quiz	104
C Participant Feedback	108

List of Figures

1.1	Comparison of NDCG@5 Calculations for Algorithm A and Algorithm B	4
3.1	Recruitment text shown to MovieLens visitors during the study	28
3.2	Sample Movie Judgement Page from Study Interface	36
3.3	Sample Movie Ranking Page from Study Interface	37
3.4	Funnel Plot of Study Stages	41
4.1	Distribution of Participant Age	45
4.2	Distribution of Number of Ratings per Person - MovieLens Users (top) and Selected Participants (bottom)	49
4.3	Average Rating Distribution per User	51
4.4	Distribution of of Average Rating Distribution - MovieLens Users (top) and Selected Participants (bottom)	54
4.5	Distribution of Last Year of Movie Rated - MovieLens Users	55
4.6	Distribution of Release Year of Rated Movies - MovieLens Users (top) and Selected Participants (bottom)	57
4.7	Distribution of Lifespans - MovieLens Users (top) and Selected Participants (bottom)	59

4.8	Difference in Interest Levels between Seen and Unseen Movies	72
4.9	Distribution of Interest Levels by Movie Popularity Zone	77
4.10	Distribution of Movie Years by Interest Level	78
4.11	Importance of Features in Interest Level Prediction	81
A.1	First part of the overview page	94
A.2	Second part of the overview page	95
A.3	Login page	95
A.4	Home Page	96
A.5	Default view of the movie judgement page	98
A.6	List of judged movies	99
A.7	Judge completion page	100
A.8	Movie ranking page	101
A.9	Ranking page unavailable	102
A.10	Post study questionnaire	102
A.11	Final web application page	103

List of Tables

3.1	Epoch limit and evaluation step for each algorithm	24
3.2	Optimal Number of Training Epochs for Each Algorithm	24
3.3	Item-based k-NN Instances	27
3.4	User-based k-NN Instances	27
3.5	Funk SVD Instance	28
3.6	BiasedMF Instance	28
3.7	ImplicitMF Instance	29
3.8	Used Dataset Versions By Each Algorithm	33
4.1	Age Statistics for Selected Participants and ML-1M	44
4.2	Gender Statistics for Selected Participants and ML-1M	44
4.3	Ethnicity Statistics for Selected Participants	46
4.4	Education Statistics for Selected Participants	47
4.5	Number of Ratings Per User for MovieLens Users and Selected Participants	50
4.6	Distribution of Rating Values for MovieLens Users and Selected Participants	52
4.7	Distribution and Count of Last Year of Movie Rated for Selected Participants	53

4.8	Year Statistics of Rated Movies for MovieLens Users and Selected Participants	56
4.9	Summary of User Lifespan	58
4.10	Test-Retest Reliability Statistics	61
4.11	Statistics of Consistency Item Past Views	63
4.12	Statistics of Number of Judgements and Pooling Depths	65
4.13	Interest Levels vs. Predicted Rating for Unseen Movies (%)	67
4.15	Interest Levels vs. Predicted Rating for Seen, Unrated Movies (%)	68
4.17	Interest Levels vs. Predicted Rating for Seen, Consistency Check Movies (%)	70
4.19	Interest Levels vs. Predicted Rating for All Seen Movies (%)	71
4.21	Interest Levels vs. Familiarity (%)	74
4.23	Summary Statistics of Average Judgement Time (in seconds)	79

Chapter 1

Introduction

Recommender systems have become an important part of our digital experiences, particularly in the domain of movie watching. These systems help users go through vast amounts of content to find films they're likely to enjoy. A recommender system analyzes user preferences and actions to suggest relevant items, which in this thesis are movies. As these systems are continuously evolving, the methods used to evaluate their performance need to improve as well. This thesis presents a novel approach to recommender system evaluation that is inspired by information retrieval test collection construction techniques.

1.1 Research Motivation

Traditionally, recommender systems were usually evaluated by using methods from machine learning by using historical ratings from users to assess algorithm performance (Cañamares et al., 2020). The process involved splitting the dataset into training and test sets. The training data was utilized to build and fine-tune the algorithm. The test set contains unseen data and was used to evaluate the algorithm's predictive accuracy. The way a

system or algorithm is evaluated is to check its ability to predict accurate ratings based on the test set.

Root Mean Square Error (RMSE), which measures the average magnitude of prediction errors, was mostly used to assess rating accuracy. The recent shift in recommender systems from predicting rating accuracy to evaluating ranking quality of lists with metrics like Normalized Discounted Cumulative Gain (NDCG) was a step forward (Ekstrand et al., 2011). Although these approaches have been widely used, there are some limitations that reduce their effectiveness for real-world scenarios.

The first problem with this traditional method is the fact that users only provide rating for a subset of movies they have seen. This becomes an issue for evaluation measures, as they only have access to data on rated items, while unrated movies might also have been seen before by a user. Moreover, the ratings provided by users do not have a random distribution, as the pool of items that users have previously rated are influenced by other recommendation algorithms or other factors as mentioned by Marlin et al. (2012): For instance, popular items have a higher chance to be rated while niche movies receive far fewer interactions from users. Also, if no one has rated a movie, it is most likely not going to be recommended by an existing system, which leads to others not being exposed to it, and the loop continues. Consequently, many discrepancies will happen, as many recommended items may end up having no or very few user ratings. This leaves no way to assess their recommendation quality with existing methods.

This setting can lead to a situation in which different lists of recommendations are tagged as equally good. For better understanding, consider a case where we have a random user who hates horror movies and therefore, has not provided any rating for this genre in their profile. Imagine we also have two algorithms named A and B at hand. They each provide a top five movie recommendation list. Both lists have a highly-rated movie (e.g., Star Wars) in the fifth position. Assume that we know the user has previously provided a

rating of 5 for that item. However, the top four recommendations from algorithm A include one horror science fiction movie, e.g., Alien. On the other hand, algorithm B recommends romance and drama films as the top four. If the user hasn't rated any of these top four recommendations, then no interaction between this user and those movies exists in the test or training set. Now, let's compare these two recommendation lists. NDCG is a commonly used metric for evaluating ranking systems. It measures the quality of a ranking list based on the position of relevant items. The basic formula for NDCG at position k is:

$$\text{NDCG@k} = \frac{\text{DCG@k}}{\text{IDCG@k}}$$

Where DCG@k is the Discounted Cumulative Gain up to position k . The discounting is used to give less weight to relevant items that appear at the end of the list. If we consider a relevance score ($rel(i)$) for each item that is at position i , the formula for DCG@k would be:

$$\text{DCG@K} = \sum_{k=1}^K \frac{rel_i}{\log_2(i+1)}$$

and IDCG@k is the best possible DCG up to position k . Ideal DCG comes from having a sorted list of items based on their relevance score from top to bottom. When there is no previous rating for an item, there is no way to know its relevance score using this evaluation metric, therefore, the score for these items is usually zero. Let's assume that the relevance score of each item is its previous rating from that user, so the higher the rating, the more relevant the item. In our example, when calculating NDCG@5 for both recommendation lists, the results would be identical because the only rated item in the test set (Star Wars with a rating of 5) appears in the same position (5th) in both lists. This movie was included in the recommendations because it was part of the test set, for which the recommendation algorithms didn't have access to the user's ratings. The top 4 items in both lists have no ratings in the test set, so they don't contribute to the NDCG score. Figure 1.1 illustrates this setting.

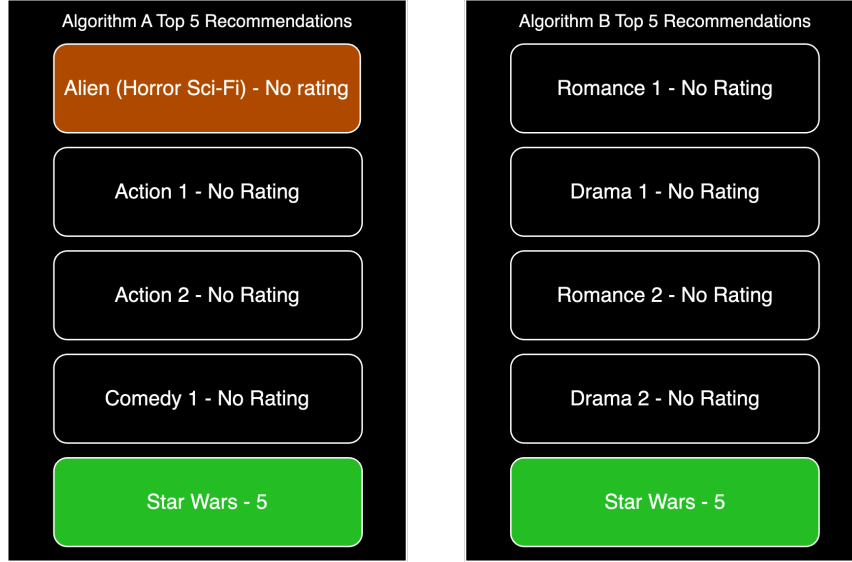


Figure 1.1: Comparison of NDCG@5 Calculations for Algorithm A and Algorithm B

Since the test set is the same for both algorithms, their IDCG values would be the same and we only need to see DCG@5 for both lists:

$$\text{DCG@5} = 0 + 0 + 0 + 0 + \frac{5}{\log_2(5+1)} = 1.93$$

$$\text{NDCG@5} = \frac{\text{DCG@5}}{\text{IDCG@5}} = 1.93/5.0 \approx 0.38$$

Therefore, these two systems are considered equally effective. However, the user knows that they dislike any kind of horror movies. Therefore, any list of recommendations that includes horror movies is considered poor recommendation for them. But the issue here is that our evaluation of these two different recommendation algorithms is flawed without knowledge of what the user thinks of these movie recommendations. Assuming that the user would hate to be recommended Alien and would not mind the other recommendations, we'd be able to say that Algorithm A is providing worse recommendations than Algorithm

B, but under the traditional test/train split methodology used in recommendation systems evaluation, the algorithms look the same, but they are very different.

To address these limitations, we draw inspiration from information retrieval (IR) test collection methods. In this setting, we have some documents, search topics, and relevance judgments for each topic that indicate the relevance of documents for that topic. We view search topics as users. Therefore, instead of focusing on predicting ratings for consumed items, we aim to evaluate a system’s ability to recommend items that users are interested in consuming.

After recruiting participants, we utilized their full profile of rated movies as input for a diverse list of recommendation algorithms. Then, we pooled the outputs of each algorithm for each participant. Pooling is a method where one combines (or “pools”) the top-ranked results from multiple recommendation algorithms to create a single set. The next step was having participants judge these recommendations based on their interest in watching rather than their satisfaction after watching. So we move away from “what is your rating after watching this item” to “how interested are you in watching this item”.

This approach allows us to collect user relevance judgments for the top recommended movies, thus minimizing the number of recommended movies for which we don’t know if the user would like to be recommended them or not. By doing so, we address one of the key limitations of traditional offline evaluation methods, where many recommended items lack user feedback.

This process resulted in the creation of a novel test collection from 51 participants with 123104 ratings and 31236 relevance judgments.

1.2 Thesis Overview

In this research, we built a test collection using standard techniques from information retrieval by recruiting Movielens users directly from its website. For each phase, we provided each participant with personalized recommendations from a diverse list of both traditional and state-of-the-art algorithms, and gathered their feedback on multiple factors such as familiarity, previous or perceived rating, and their interest in watching each recommended movie. Finally, participants were asked to provide a ranking of their top three titles from the pool of their most desired movies. [Chapter 2](#) is focused on previous work in the field, especially test collections in information retrieval, various modes of evaluating recommender systems, and the history of Movielens datasets. [Chapter 3](#) explains the process of recruiting participants, gathering rating profiles from them, and training multiple models and compiling the pool of recommended items to be shown to each user. In [Chapter 4](#), we report our various findings from the user study including participant demographics, their differences with average Movielens users, and the factors that could describe their willingness to watch an item. Lastly, [Chapter 5](#) wraps up this thesis.

1.3 Contributions

This thesis makes the following contributions:

- Constructing a novel test collection for recommender system evaluation.
- Providing a detailed analysis of user behavior and insights into the relationship between predicted ratings, familiarity, and interest in watching movies.
- Identifying significant differences between highly engaged MovieLens users (study participants) and average users.

Chapter 2

Related Work

2.1 Test Collections in Information Retrieval

As defined by [Sanderson et al. \(2010\)](#), a test collection in information retrieval (IR) is made up of some documents, queries or search topics, relevance judgements. The aim of having this definition and setting is to find out which documents are relevant with respect to each query. One can also view test collections as datasets. Using a standardized test collection, various algorithms and retrieval methods can be evaluated and compared under consistent conditions ([Buttcher et al., 2016](#)).

It is common to credit the origins of modern retrieval system evaluation to the 1960s Cranfield experiments. These experiments, which were first conducted by [Cleverdon \(1967\)](#), marked an important change in the direction of controlled, empirical testing. Their focus was on the effectiveness of various indexing languages and how to improve the overall retrieval quality of IR systems.

On the other hand, efforts have been made to establish ongoing workshops with the goal of encouraging the use of suitable evaluation measures and the development of innovative

evaluation methods. The most prominent workshop in the domain of IR is the Text REtrieval Conference (TREC) (Harman and Voorhees, 2006). The conferences are held by the National Institute of Standards and Technology (NIST) of the US government and the first one was in 1992 (Voorhees et al., 2005). One of the many benefits of TREC evaluation approaches has been the enhancement of numerous search methods. According to Tassey et al. (2010), from 1999 up to 2009, because of the presence of TREC, users based out of US saved about 3.15 billion hours in internet usage time.

2.2 Evaluation in Information Retrieval

There are two main methods to evaluate IR systems, namely online and offline evaluation methods. Online evaluation revolves around assessing the quality of an IR system in real-time with actual users. As Hofmann et al. (2016) described, this method looks into how users work with a live system, and it tracks metrics such as session time and click-through rates. The advantage of this approach is that it provides insights about real user behavior. However, the downside is that it can be expensive and time-consuming. Whereas offline evaluation relies on test collections, as mentioned earlier. It enables assessing IR systems without the need for actual users (Voorhees, 2001). This setting makes this approach highly reproducible and efficient.

2.3 Recommender Systems

Nowadays, we can observe the presence of recommender systems in almost every aspect of our digital lives. These systems and their numerous and diverse algorithms shape many of our daily decisions (Lü et al., 2012). As indicated by an overview from Burke et al. (2011), several artificial intelligence (AI) approaches, including but not limited to “machine learn-

ing (ML), data mining, user modeling, and constraint satisfaction”, have been employed within recommender systems research.

With the importance and prevalence of these systems in mind, the question of how to properly evaluate the performance of systems becomes essential. Many studies have been carried out in regards to categorizing the various experiment types one can choose for the evaluation phase. As a result, we can define three different types of experiments: online evaluation, user study, and offline evaluation (Beel and Langer, 2015; Gunawardana et al., 2012). A combination of experiment types is required for a full understanding of a recommendation system’s performance. In support of this claim, Freyne and Berkovsky (2013) stated that although the first stage is to make use of offline evaluations, it is necessary to progress to user studies and conclude with online assessment in order to ensure comprehensive evaluation coverage.

2.3.1 Online Evaluation

Most online evaluations have happened as A/B tests in which a baseline system is compared against the newly created system(s), as pointed out by Kohavi et al. (2013). The primary aim of this kind of evaluation is that the recommendation system is added to a tool that is already available and user-facing (i.e, a live website), and users keep working with the system just like before (Zangerle and Bauer, 2022). In this scenario, since users have autonomy and access the system in the most intuitive and real way feasible, online evaluation provides a realistic assessment (Kohavi et al., 2009). Moreover, this kind of experiment is necessary when researchers are evaluating users interactions with the system and can be applied to almost every recommender system research venue.

2.3.2 Offline Evaluation

Based on the findings from [Jannach and Bauer \(2020\)](#), of the 117 recommendation system articles published in 2018 and 2019 at AAAI and IJCAI, almost 92% relied solely on offline experiments. Therefore, this type of experiment is believed to be the most widely used method for evaluation. This evaluation technique utilizes previously collected datasets. In these evaluations, by using the available historical data, the behavior of users is replicated through simulations. The main difference to online evaluation is that no real users interacting directly with the system during the experiments are available ([Gunawardana et al., 2012](#)). Unlike user studies and online evaluations, the focus here is to compare various recommender system algorithms against each other in terms of accuracy and precision in predicting user preferences ([Beel et al., 2013](#)). As indicated by [Shani and Gunawardana \(2011\)](#) and based on the characteristics described, offline evaluations are highly repeatable since each run can be customized with various configurations of setups, algorithm parameters, and dataset portions.

[Castells and Moffat \(2022\)](#) highlights that offline evaluation in recommender systems has challenges compared to traditional machine learning, or IR. Unlike IR, where relevance judgments can be made independently of users, recommender systems are dependent on individual user preferences. Therefore, the ground truth for evaluating recommendations is difficult to get at scale. However, it is important to note that our study directly tackles this challenge by creating a test collection for offline evaluation of recommender systems. Another interesting area the authors point out is that offline evaluation aims to predict what happens during online performance, but the correlation between offline and online results is often weak.

Despite its widespread use, offline evaluation is not without its issues. [Hidasi and Czapp \(2023\)](#) identified several common problems in offline evaluation preparation. These include mismatches between dataset and task in which datasets are used for some tasks they're

not suitable for, unnecessary preprocessing that can lead to less generalizable results, and information leaking through time because of incorrect train-test splits.

2.4 Advancements in Offline Evaluation Methodology

As detailed in section 2.2, in IR, online evaluation refers to assessing a running system based on real user behavior, while offline evaluation uses existing test collections without the need for user interaction. However, in recommender systems research, these terms have sometimes been used differently, potentially leading to confusion.

A notable study by Rossetti et al. (2016) attempted to contrast what they termed “offline” and “online” evaluation methods for recommender systems. In their study, they used the MovieLens dataset to train four different recommendation algorithms: Most-popular, Matrix Factorization with 80 factors, Matrix Factorization with 400 factors, and Item-based nearest neighbor. For their online evaluation, they invited participants to rate movies and then evaluate recommendations generated by these algorithms. They found that the ranking of algorithms from offline measurements were not at par with the results of online user study. Furthermore, they reported that the utility of recommendations, defined as previously unknown items that users liked, produced a totally different ranking of algorithms compared to both offline and online evaluations.

However, their approach is different from the traditional IR understanding of these concepts. Their “offline” evaluation is a standard train-test split methodology, while their “online” evaluation involved direct user feedback on recommendations—a process closer to what we consider offline evaluation in IR. Our study builds upon and expands this earlier work in several important areas, which are as follows:

- **Data:** We used the ML-32M dataset which was updated shortly before participant recruitment, compared to the older and smaller ML-1M utilized in their study.

- **User profiles:** Instead of asking participants to select movies they like, we considered complete and real MovieLens user profiles to have an authentic representation of user preferences.
- **Evaluation Depth:** Our study evaluated a much larger pool of recommendations per participant (up to 600 and average depth of 60 for ranked recommendation lists) compared to the maximum of 20 movies they recommended to each user.
- **Algorithm diversity:** We included a wider and more diverse range of recommendation algorithms (11 categories and 22 unique algorithms) in our evaluation (compared to 4 algorithms they had).
- **Feedback:** While both studies asked similar questions about familiarity with recommendations, we collected more detailed feedback that included ratings for watching movies and a 5-point scale for interest in watching.
- **Test Collection:** Unlike the previous study, our work aims to produce a reusable test collection for the research community.

Our process supports and extends the conclusions of [Rossetti et al. \(2016\)](#) and highlights the significant differences between traditional split-based evaluation and methods that consider actual user ratings and opinions.

2.5 MovieLens Datasets

The first version of the MovieLens dataset was released back in 1998 by the GroupLens research group from the University of Minnesota. From then, multiple versions of this dataset have been made available and have played an important role in recommendation systems research ([Harper and Konstan, 2015](#)). Their research group also maintains the

MovieLens website¹. The website has the same functionalities as IMDB, such as hosting up-to-date information about released movie titles. Website users can also provide ratings for movies in the form of stars (on a scale of 0.5 to 5). The MovieLens datasets are compiled from user interactions on the website.

Modifications to the website’s appearance and features, such as half-star ratings and improved search abilities, have had a big impact on the growth and development of the MovieLens datasets. As reported, these datasets were downloaded more than 140,000 times in 2014 alone and were heavily referenced in academic work (approximately 2,075 registered citations as of this writing).

Furthermore, the datasets have made a wide range of studies and experiments possible, which has advanced customization of recommender systems. They have been particularly vital in exploring various algorithmic developments, such as collaborative filtering and dimensionality reduction techniques (Alam et al., 2021).

2.5.1 MovieLens Versions and Use Cases

As previously mentioned, there are multiple versions of the MovieLens dataset, and each one is targeted for a specific scale of research. All of them are available at GroupLens². The datasets are named after the number of interactions they have. For instance, ML-100K contains 100,000 movie ratings (interactions) and is usually used in introductory data science and ML courses due to its manageable size. On the other hand, released in December 2019, and one of the latest in this series is ML-25M, which features 25 million ratings for 62,000 movies by 162,000 users. The larger datasets are mainly considered for new research.

¹<https://movielens.org>

²<https://grouplens.org/datasets/movielens/>

Chapter 3

Creating the Test Collection

Before recruiting participants for the user study, we needed to implement a data cleaning process, a recommendation pipeline, and a custom website with detailed instructions. This chapter describes the steps taken to create the test collection.

3.1 Dataset Preparation

For the base dataset, we utilized the most recent version of the MovieLens dataset available on September 2023 (Harper and Konstan, 2015), that was publicly accessible at <https://files.grouplens.org/datasets/movielens/ml-32m>. From now on referred to as ML-32M in this thesis, this dataset is made up of 32,000,204 ratings and 84,432 movies from capturing the interactions of 200,948 users with the movielens.org website.

Moreover, for use in some parts of model fine-tuning (later described in subsection 3.2.3) and user demographic analysis, we also utilized another version of the MovieLens dataset called ML-1M. This dataset is made up of 1,000,209 ratings applied to a pool of 3,900 movie titles. There are 6,040 unique and anonymous users in this dataset.

3.1.1 Dataset Extension with Movie Details

To provide participants with comprehensive information about each movie for when they were going to provide judgments, additional data beyond the release date and title was required. For this target, we utilized the TMDB API ¹. The ML-32M dataset includes a unique TMDB identifier, “tmdbId”, for each movie, which was used to get the additional details. Also, any movie with an adult tag returned from the API was excluded from our dataset. The following is a list of added fields that are included with every movie:

- Overview: A brief summary of the movie plot and events.
- Image URL: The public URL for the movie poster.
- Runtime: Total duration of the movie in minutes.
- Cast: Names of the top five cast members.
- Directors: Names of the directors.
- Languages: Languages spoken inside the film.

3.1.2 Base Dataset Filtering

We first start by describing a filtering method in recommender systems. To perform N-core filtering on a dataset simply involves retaining all the users or items that have at least appeared in the dataset for ‘n’ times. This method is aimed at reducing dataset sparsity. This in turn improves the training process and recommendation quality. Our data preprocessing began with a two-stage filtering process. At first, we applied a 10-core filtering to movies. In other words, we removed those movies that had less than 10 appearances.

¹<https://developer.themoviedb.org/docs/getting-started>

Second, based on the extra information that we had available as a result of extending dataset with movie details (subsection 3.1.1), any items tagged as TV shows or series were excluded from the dataset. Afterwards, with the help of Abdul Manaam, we manually checked the remaining movies that had an invalid or incorrect TMDB identifier based on the data retrieved from the TMDB API. If any item belonged to a TV show or series, it was removed. If it belonged to a valid movie, we found and replaced the correct TMDB identifier for it. Lastly, we eliminated all user interactions with the movies that were removed in the previous steps.

Following the movie filtering, we conducted a 20-core filtering on users to make sure that each user had interacted with at least 20 movies. Lastly, another round of 10-core filtering on movies was performed to discard any titles that, after user-based filtering, appeared fewer than 10 times in the dataset. After these steps, the dataset had 31,696,218 ratings from 200,797 unique users applied to 31,271 distinctive movie titles. This was the base version of the dataset used in this study. For each one of the three batches of participants (described in Section 3.4), we also performed an extra round of filtering that is fully explained in Subsection 3.4.2.

3.2 Recommendation Algorithms

After the process of pre-cleaning, extending, and filtering the base dataset, we continued to prepare the recommendation algorithms. The effectiveness of a recommendation system is highly dependent on its choice of algorithms. In this section, we present an overview of all the recommendation algorithms used, along with the process of fine-tuning and obtaining their best hyper-parameter values.

3.2.1 Recommendation Frameworks

Two recommendation system frameworks were utilized in this study: LensKit and RecBole. Before moving on to the algorithm interface, we first present a brief overview of these two frameworks. LensKit is an open-source framework upgraded from its original Java-based implementation introduced in 2010 to a Python package (Ekstrand, 2020).

The second framework we used to widen the range of models and also include more state-of-the-art methods was RecBole (Zhao et al., 2022). We first start by providing a detailed description of each one of the chosen algorithms alongside the final values for their hyperparameters.

3.2.2 Algorithm interface

We employed a diverse set of recommendation algorithms that include simple baseline methods, traditional and state-of-the-art collaborative filtering, and more advanced neural network-based approaches. The algorithms used in our experiments are:

- (1) Bias-based algorithms:
 - **Global Bias - LensKit**: An algorithm that aims to predict the ratings based on the global average rating. It then adjusts this value for specific user and item biases. Its aim is to follow the general tendencies in data.
 - **Popular - LensKit**: It recommends the most popular items, based on their number of ratings. There is no consideration of different user tastes in this algorithm. There are also no parameter configurations for it.
- (2) Neighborhood-based methods:

- **Item-kNN - LensKit**: Recommends items similar to those the user has liked in the past. It uses cosine similarity to build an item-item matrix and uses it to make predictions ([Deshpande and Karypis, 2004](#)). There are two versions of this algorithm that work with either implicit or explicit forms of feedback.
- **User-kNN - LensKit**: Groups users into similar sets according to their observed rating history. The main task is then to identify said groups of users. We refer to them as “neighbors”. Neighbors are the ones whose taste is similar to the current user. Like Item-kNN, this algorithm has two versions that can consume either implicit or explicit feedback.
- (3) Matrix Factorization Techniques:
 - **FunkSVD - LensKit**: First introduced by [Funk \(2006\)](#), FunkSVD uses stochastic gradient descent to optimize user-feature and item-feature matrices.
 - **BiasedMF - LensKit**: Introduced by [Zhou et al. \(2008\)](#), this algorithm essentially lowers the mean squared error (MSE) between observed ratings and model predictions by alternately fixing one matrix and optimizing the other.
 - **ImplicitMF - LensKit**: Similar to BiasedMF but used for implicit feedback and developed by [Hu et al. \(2008\)](#).
- (4) Advanced collaborative filtering methods:
 - **ADMM SLIM - RecBole**: It was introduced by [Steck et al. \(2020\)](#) and is an advanced version of the Sparse Linear Method (SLIM) that uses the Alternating Directions Method of Multipliers (ADMM) to efficiently handle large and sparse datasets.
 - **BPR - RecBole**: Bayesian Personalized Ranking (BPR) is a ML approach for personalized ranking of unseen items ([Rendle et al., 2012](#)). BPR optimizes a

specific feature called BPR-Opt and was created to work with implicit feedback as input.

- **CDAE - RecBole**: Collaborative Denoising Auto-Encoders (CDAE) utilize the framework of Denoising Auto-Encoders to optimize the top-N recommendation task and effectively predict user preferences. There is a unique feature in this algorithm that allows the model to learn and recover user preferences more robustly (Wu et al., 2016).
- **EASE - RecBole**: Embarrassingly Shallow Autoencoders for Sparse Data (EASE) by Steck (2019) introduces a linear model that is specifically designed for situations in which the data is sparse and in implicit feedback form. It functions without any hidden layers and predicts item recommendations for users directly by simulating input interactions as outputs.
- **MultiVAE - RecBole**: MultiVAE is a variational autoencoder and like EASE, it is aimed to be used with implicit feedback data (Liang et al., 2018). It provides a proxy for the top-N ranking loss by using a multinomial likelihood.
- **NeuMF - RecBole**: First introduced by He et al. (2017), NeuMF is an implementation of the Neural Collaborative Filtering (NCF) method. In addition to NCF, NeuMF combines matrix factorization with multi layer perceptrons to increase the quality of recommendations.

3.2.3 Fine-Tuning RecBole Algorithms

The next step was to find the best set of parameters for each of the chosen algorithms. This process is referred to as hyper-parameter tuning. Before going into the specifics of our tuning process, it's important to cover two common ways to hyper-parameter optimization: grid search and sequential search. Grid search is a comprehensive search method

that considers all possible combinations of chosen parameters and their values. Although it evaluates the performance of all potential models, it can be expensive in terms of computation. On the other hand, we have sequential search. It is a more efficient alternative to grid search. This method considers the result of previous iterations to select the next combination of parameter values. Therefore, it focuses on the most promising areas in the parameter space and reduces the number of needed evaluations.

In our parameter tuning process, we had two approaches for algorithms from RecBole versus those of LensKit. We start by describing the process for RecBole models.

ADMM SLIM

There are three main hyperparameters in the RecBole implementation of this model:

- λ_1 and λ_2 , which represent the L1 and L2 norm regularization parameters.
- α , which controls the power-law of the regularization terms.

To identify the optimal hyperparameters, we performed a sequential search on the ML-1M dataset and best hyperparameter settings were found to be:

```
alpha: 1
lambda1: 5
lambda2: 1000
```

BPR

To find the optimized values for these two parameters, we turned to literature. [Zhao et al. \(2022\)](#) performed a comprehensive grid search on the ML-1M dataset for a multitude of algorithms. The optimal parameter values for the BPR model were selected as:

```
embedding_size: 2048
learning_rate: 0.0001
```

CDAE

We utilized the parameter values found by [Zhao et al. \(2022\)](#) for CDAE. The authors performed a sequential search with an early stopping. Additionally, we made adjustments to the default L1 and L2 regularization weights during the process. The final values were as follows:

```
embedding_size: 64
learning_rate: 0.05
reg_weight_1: 0.01
reg_weight_2: 0
```

EASE

EASE only had one hyperparameter to be tuned which is the L2 regularization weight. Similar to ADMM SLIM, the optimal value for this parameter was obtained through a sequential search over the ML-1M dataset. The regularization weight that provided the best result was found to be 500.

MultiVAE

There were a few configurable parameters for MultiVAE which are as follows:

- **MLP hidden layer sizes:** Sets the vector size of the different hidden layers inside the neural network, with the array length as the layers count.

- **Dropout Probability:** Helps reduce overfitting by randomly disabling neurons during training.
- **Anneal Cap:** Limits the influence of the KL divergence in the early stages of training and allows for more stable convergence ([Bowman et al., 2015](#)).

The final values for the parameters were obtained using the sequential search process over the ML-1M dataset:

```
mlp_hidden_size: [300]
dropout_prob: 0.3
anneal_cap: 0.1
learning_rate: 0.01
```

NeuMF

The same sequential search was performed for NeuMF and the final settings determined for the model were as follows:

```
mlp_hidden_size: [256, 128, 256]
dropout_prob: 0.2
learning_rate: 0.0005
```

Epoch Choices

It is customary to train recommendation algorithms for multiple rounds and keep improving the algorithm’s ability to capture user tastes by updating the model. The number of training rounds is usually denoted by epochs. To set the number of training epochs for each one of the considered algorithms, we utilized the evaluation module provided by

the RecBole framework (Zhao et al., 2020). For this task, we used the output of the data filtering step (subsection 3.1.2) which was the cleaned version of ML-32M. We used a random ordering and ratio-based splitting technique to divide the data. Random ordering means users and their data are shuffled randomly to mitigate any ordering bias. This setting meant allocating 90% of each user’s ratings to the training set and 10% to the development (validation) set. The metric used to evaluate training effectiveness and for early stopping of the process was the mean average precision at depth 1000 (MAP@1000), which is defined as follows:

$$\text{MAP@1000} = \frac{1}{|U|} \sum_{u=1}^{|U|} \sum_{k=1}^{\min(t, 1000)} P(k) \times \text{rel}(k)$$

where t is the number of relevant items for the user, $|U|$ is the number of users, $P(k)$ is precision at depth k in the ranked list, and $\text{rel}(k)$ is a custom binary function that outputs 0 if the item at rank k is considered non-relevant and 1 otherwise.

To improve the search process and make sure it was finished on time, three extra configuration parameters were set for every algorithm:

- *epochs*: Specifies the maximum number of training epochs.
- *eval_step*: Determines how frequently to perform evaluation on the validation set. A value of 3 means doing one evaluation after every three epochs.
- *stopping_step*: The number of consecutive evaluation steps without improvement after which training process is stopped to prevent overfitting.

The values for these parameters are summarized in Table 3.1.

These settings were combined with the individual model parameters previously described to find the best number of training epochs for each algorithm with MAP@1000 as the deciding factor. Table 3.2 shows the final number of epochs for each model.

Table 3.1: Epoch limit and evaluation step for each algorithm

Model	<i>epochs</i>	<i>eval_step</i>	<i>stopping_step</i>
ADMM SLIM	500	1	10
BPR	1000	10	5
CDAE	500	3	10
EASE	500	1	10
MultiVAE	500	3	10
NeuMF	500	3	10

Table 3.2: Optimal Number of Training Epochs for Each Algorithm

Model	Optimal Epochs
BPR	659
NeuMF	137
CDAE	176
MultiVAE	227
EASE	1
ADMM SLIM	1

3.2.4 Fine-Tuning LensKit Algorithms

We now describe the process followed for the algorithms from the LensKit framework.

Parameter Introduction

As for the LensKit algorithms, before delving deeper into the process of their fine-tuning, we first provide a description of the parameters frequently used to customize this framework’s models. Note that in LensKit, the terms “score” and “rating” are used interchangeably.

- **nnbrs**: The maximum count of neighbors to choose for rating prediction.
- **min_nbrs**: The minimum count of neighbors to choose for rating prediction.
- **min_sim**: Minimum similarity threshold to consider another user as a neighbor.
- **feedback**: How user feedback is interpreted, either as explicit or implicit.
- **features**: The number of features to be used in matrix factorization models.
- **damping**: The damping factor applied to the interaction matrix of users and items.
- **iterations**: The number of training epochs to run for each feature.
- **lr**: The learning rate to control the speed of model parameter updates.
- **reg**: The regularization factor to avoid model overfitting.

Parameter Selection Process

Our parameter selection process for LensKit algorithms had several steps to reach optimal performance. We began by utilizing the hyperparameter optimization function from the LensKit auto tool (Vente et al., 2023) on the ML-1M dataset. This first step gave us baseline values for each LensKit algorithm’s parameters. After that, we followed out these steps:

Professor Smucker performed a demo run using an earlier version of the website and the process described in section 3.4 in which he judged his own pool of recommendations based on the outputs of LensKit algorithms with their parameters obtained from LensKit auto tool.

We then did an extensive exploration of the parameter space using both Professor Smucker’s judgments and a sample of 1,000 users from the filtered version of ML-32M

(subsection 3.1.2). For the sampled users we used a 75 to 25 train/test split and evaluated the quality of recommendations using nDCG at depth 1000 (nDCG@1000). For Professor Smucker’s recommendations, we considered both nDCG@1000 and compatibility with persistence probabilities of 0.95 and 0.98.

Furthermore, for each algorithm, we considered the parameter settings suggested by Michael Ekstrand that he personally communicated with us.

Next, we analyzed which parameter settings performed well for the 1,000 sampled users with traditional evaluation techniques, and which settings worked well for Professor Smucker’s profile using a setup similar to our planned test collection evaluation. These two criteria were the baseline of our selection process for the different combinations of parameter values for each algorithm. According to this process, we now present the final configuration of parameters for each LensKit algorithm.

Bias

We set the damping factor of this algorithm to 5 as it was the only configurable parameter.

Item-kNN

Table 3.3 demonstrates the various instances of Item-kNN we made for this study, alongside their final parameter values.

User-kNNs

As before, table 3.4 depicts instances of User-kNNs and their parameters considered.

Table 3.3: Item-based k-NN Instances

Instance Name	feedback	nnbrs	min_nbrs	min_sim
IIEx_30_10_005	Explicit	30	10	0.05
IIEx_30_2_001	Explicit	30	2	0.01
IIEx_30_30_005	Explicit	30	30	0.05
IIIm_1_1_0001	Implicit	1	1	0.001
IIIm_120_15_0001	Implicit	120	15	0.001

Table 3.4: User-based k-NN Instances

Instance Name	feedback	nnbrs	min_nbrs	min_sim
UUEx_120_2_001	Explicit	120	2	0.01
UUEx_120_30_001	Explicit	120	30	0.01
UUEx_30_2_01	Explicit	30	2	0.1
UUEx_30_30_01	Explicit	30	30	0.1
UUEx_60_20_0075	Explicit	60	20	0.075
UUIIm_30_2_001	Implicit	30	2	0.01

FunkSVD

Table 3.5 demonstrates the parameter settings for the only FunkSVD instance in this study.

BiasedMF & ImplicitMF

For the instance of BiasedMF, we set the number of features to 250 and also found that the default number of iterations (20) was the best. Table 3.6 showcases these values.

Just like BiasedMF, we changed the number of latent features to 250 and kept the other settings to their default, as shown in table 3.7.

Table 3.5: Funk SVD Instance

Instance Name	damping	features	iterations	lrate	reg
FunkSVD	5	250	175	0.001	0.015

Table 3.6: BiasedMF Instance

Instance Name	damping	features	iterations	reg
BiasedMF	5	250	20	0.1

3.3 Participant Recruitment

After discussing the research goals with the GroupLens research group at the University of Minnesota, they agreed to support our study by putting up the recruitment banner on the movielens.org website. This banner was shown to every MovieLens user at the top of the homepage from October 17, 2023, until December 6, 2023. Figure 3.1 shows the text of this banner.

Banner Text shown on MovieLens:

University of Waterloo researchers are looking for MovieLens users to participate in a study that measures recommendation system quality. Participants will get an Amazon e-gift card worth \$15 US or \$20 CAD per hour (2-5 hours of time commitment). Click here for more information. *[This study has been reviewed and received clearance through a University of Waterloo Research Ethics Board]*

And by clicking on the link, users will be redirected to a webpage where they can read full details of the study, fill out the screening questionnaire, and hit submit.

Figure 3.1: Recruitment text shown to MovieLens visitors during the study

If MovieLens users were interested, users could click on a link in the banner text that would redirect them to our website. The page they were first shown, which we call the

Table 3.7: ImplicitMF Instance

Instance Name	features	iterations	reg
ImplicitMF	250	20	0.1

“overview” page, included a detailed overview of our research, its goals, and the study process.

After that, those individuals who were interested in participating were asked to sign up for the study by providing their full name, email address, age, and their country of residence. Based on our ethics application, we were only looking to recruit participants from the USA or Canada. This is why we asked for information on residence. The overview page also conveyed to visitors that providing their information was just an expression of interest in the study and not an obligation to participate or an expression of consent. Over the course of the study, we received 367 sign-ups from individuals all around the world.

3.4 Website Design and Study Pipeline

We present the details and step-by-step explanation of the process that participants went through after expressing interest in the study until they were finished. We also present screenshots of the user interface that was specifically created for this study.

It is important to note that our recruitment banner was displayed for an extended period, and Movielens users started signing up from the first day until the very last. We used a batch processing system to avoid delays for participants. This means that all steps mentioned in this section were carried out for three separate batches of participants. Therefore, we began the study with those who signed up earlier while still accommodating those who signed up later.

3.4.1 Screening Process and Data Obtaining

After each sign up, we analyzed two important pieces of information to determine whether each individual is eligible to participate or not. First, we verified that they were aged 18 or older, and secondly, we checked if they were residents of the USA or Canada. We informed those who did not meet the eligibility criteria that although we thank them for their time and interest, we could not use their insights in our study. For all those who passed the screening phase, we sent an email with the consent form that included the full details of each stage of the study, as well as complementary information on data security, remuneration, and the benefits of participation. It was also communicated to our participants in the consent form that they should be able to read, comprehend, and communicate in English. They were also required to have a laptop or desktop computer with a stable internet connection, as the study website was not designed to be viewed by mobile devices. Lastly, we requested that each participant return a signed copy of the consent form.

After receiving the consent form, we sent another email explaining how to export their ratings profile from the MovieLens website in comma-separated value (CSV) format. Participants were asked to send back their exported ratings in the original format. Once we received their profile, we thanked the participants for their contribution and communicated with them that the recommendation generation will shortly begin and we will get back to them with the next steps within a week.

3.4.2 Removing Potentially Existing Profiles

After that, with the filtered version of the ML-32M dataset from the output of our data filtering state (3.1.2), we looked into whether any participants of the current batch were already anonymously present in the dataset using the LensKit framework. This was a

reasonable assumption to make considering that the MovieLens datasets are sourced from actual user interactions on the MovieLens website. It was important to find and eliminate these cases because users with similar profiles could be mistakenly recognized as highly similar nearest neighbors by similarity-based algorithms.

To address this, we employed a User-kNN algorithm with parameters set to $nnbrs = 10$, $min_nnbrs = 1$, and $min_sim = 0.1$ to find the most similar existing dataset profiles for each participant. The algorithm used cosine similarity as its similarity measure and compared the rating vectors of users.

For each participant, we found the profile with the highest similarity and considered it a candidate to be match for the participant. Our code was written to output all profile matches with similarity greater than or equal to 0.9 or the sole profile with a maximum similarity below 0.9. We did not find any participant to have multiple high similarity matches, i.e. over 0.9 similarity with multiple profiles.

We manually reviewed profile matches from lowest similarity up through the matches in the 0.90-0.93 range. The matched profiles with similarity of 0.85 or greater were clearly the same person in all cases. In some cases we did consider profiles with lower scoring matches to also be the same person, but these were rare. For all three batches of participants, 103 profiles were confirmed as belonging to our participants.

3.4.3 Dataset Extension with Participant Profiles

Next, we extended the resulting dataset with the rating profiles of the batch of participants. The creation of the training dataset for the recommender system algorithms relied on this data augmentation. This is because the algorithms needed to consume information on our participants in order to come up with personalized recommendations. To make sure the final dataset was consistent, we first dropped those movies present in the participant

profile that were no longer available in the filtered dataset. During this process, there was also a check to ensure each participant has a minimum number of 20 interactions. It was important to have enough data points on each participant to be able to fully capture their preferences throughout the recommendation stage. Fortunately, every profile satisfied the requirement.

3.4.4 Explicit and Implicit Datasets

Then the first version of the final dataset was compiled. The dataset was named the “explicit dataset” since it included all the ratings from Movielens users and our participants in their original numerical form and scale, which is also known as explicit ratings.

We also needed to have an implicit dataset. According to [Jadidinejad et al. \(2019\)](#), in implicit feedback systems, the presence of interaction, such as viewing, providing feedback, or purchasing, between a user and a item indicates a preference. Therefore, in this situation, the actual value of the ratings was less important. We began creating this version of the dataset following the approach used by [Steck \(2019\)](#) in the EASE paper. This involved doing a filtering over the explicit dataset to only keep those interactions where ratings were four or greater. This process effectively binarizes the data. Next, to make sure all users have at least five or more data points, we applied a 5-core filtering to users. We named this dataset the “implicit dataset”. Another version of this dataset, named “atomic” dataset, was created to be used in the Recbole framework. The only difference between the two implicit versions were their header names. Again, it’s important to note that the “explicit”, “implicit”, and “atomic” datasets were created specifically for each batch of participants. Table [3.8](#) shows which one of these three versions of the final dataset each algorithm used.

Table 3.8: Used Dataset Versions By Each Algorithm

Dataset Version	Algorithms
Explicit	Popular, Bias, BiasedMF, FunkSVD, IIEx_30_10_005, IIEx_30_2_001, IIEx_30_30_005, UUEx_120_2_001, UUEx_120_30_001, UUEx_30_2_01, UUEx_30_30_01, UUEx_60_20_0075
Implicit	IIIm_1_1_0001, IIIm_120_15_0001, UUIIm_30_2_001, ImplicitMF
Atomic	ADMM SLIM, BPR, CDAE, EASE, MultiVAE, NeuMF

3.4.5 Recommendation Generation

With the three types of datasets (“explicit”, “implicit”, and “atomic”) at hand, during each batch of the study and from each algorithm, up to 11,000 recommendations were generated for each participant. Afterwards, we filtered out any recommended items already existing in the participants’ profiles. The results were then transformed so that only the top 1,000 recommendations are left for each participant from each model.

We then proceeded to perform a process named “pooling”. With 16 algorithms from Lenskit and 6 algorithms from RecBole, each having up to 1,000 recommendations for any single participant, it would have been nearly impossible for our participants to go through all their recommended items and provide quality judgement for each of them. Therefore, it was necessary to create a set for each one of the two stages of the study with a reasonable number of items. Here is a full description of the pooling algorithm:

1. **Perform all actions below for each participant in the same order presented**
2. **Consistency items:** Take two samples from their existing rating profile:

- (a) *p1_consistency_movieIds*: A random sample of 10 rated movies.
 - (b) *p2_consistency_movieIds*: A random sample of 50 rated movies with a different *seed* compared to the previous set.
3. Create two empty Python *sets* and name them *phaseone_unique_movieIds* and *total_unique_movieIds* respectively.
4. **Phase 1 items:**
- Loop over each recommendation algorithm’s 1,000 items
 - (a) Start from the top-ranked item (depth 1) and add the first 10 items from each algorithm’s recommendation list to both sets.
 - (b) If the total number of items in *phaseone_unique_movieIds* is less than 135:
 - i. Continue adding items from deeper in the recommendation lists until the set contains 135 unique items. Also, save the depth value.
 - (c) Merge *p1_consistency_movieIds* into *phaseone_unique_movieIds* and finalize the items to be recommended during the first study phase.
5. **Phase 2 items:**
- Starting at the last depth value from the previous step, loop over each algorithm’s output:
 - (a) While the number of unique items in *total_unique_movieIds* is less than 600 and depth has not reached 50, increment the depth value by one and add the item from that depth to *total_unique_movieIds*.
 - Merge *p2_consistency_movieIds* into *total_unique_movieIds*.
 - Compare the two sets and remove any items from *total_unique_movieIds* that is also present in *phaseone_unique_movieIds* to have the final set of items for the second phase of the study.

We then uploaded each participants two recommendation sets to our website.

3.4.6 Tutorial and Knowledge Check

As we made our research system ready for the participants to judge movies. We sent them an email with their login credentials and website URL. Upon logging into the system for the first time, each individual was shown a page with a few demographic questions for us to better understand the background of our participants. The questions asked about their gender, race or ethnic background, and highest level of obtained education. After submitting their answers, they were directed to a concise tutorial that guided them through the tasks involved in the study (Appendix A). At the end of the tutorial, there was a brief quiz made up of 12 questions to make sure that they were well-prepared to proceed (Appendix B). Once they successfully completed the tutorial and quiz, they were redirected to the study homepage to start the tasks. The homepage provided an overview of the study steps and options to begin the main tasks.

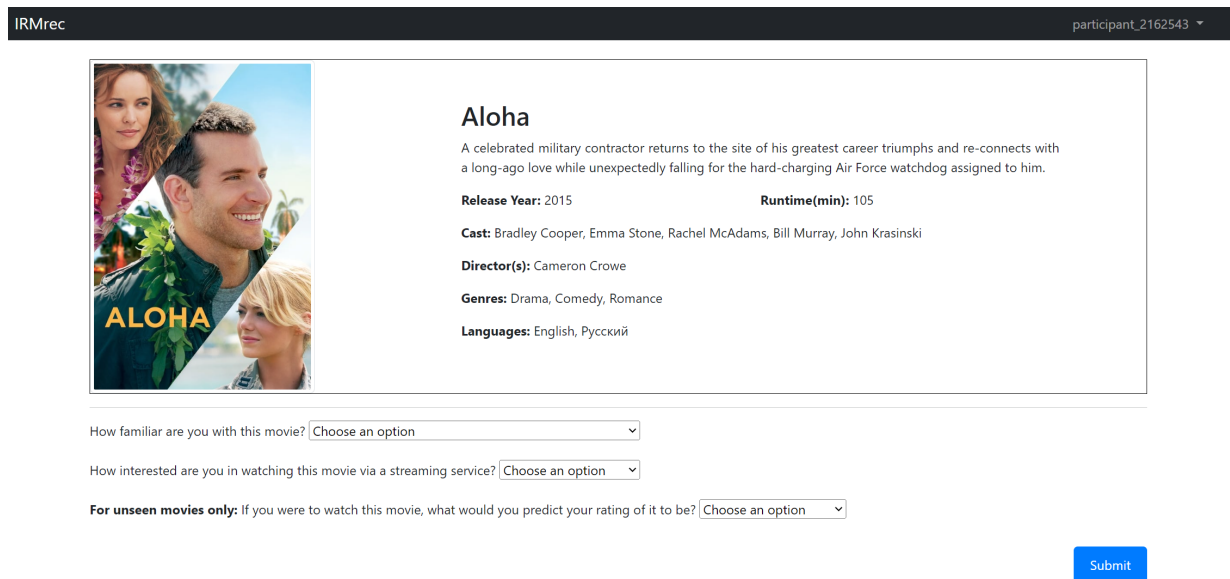
3.4.7 Phase one

The first phase of the study was designed to take up to 2 hours for each participant. The judgment task was the core part of this phase, and it involved reviewing a set of movies (approximately 150 movies). For each movie, we asked them three questions:

- **How familiar are you with this movie?** - Possible answers were: Never heard of it, Familiar with the movie, Very familiar, Have seen the movie and here is my rating (0.5 to 5 stars)
- **How interested are you in watching this movie via a streaming service?** - Possible answers were: Not interested, Somewhat interested, Interested, Very interested, Extremely interested

- **For unseen movies only: if you were to watch this movie, what would you predict your rating of it to be?** - Possible answers were 0.5 to 5 stars

Figure 3.2 shows a sample from the website interface for the three questions asked for each recommendation.



The screenshot shows the IRMrec website interface. At the top, there is a dark header with 'IRMrec' on the left and 'participant_2162543' on the right. Below the header, a movie card for 'Aloha' is displayed. The card includes a movie poster on the left and text on the right. The text includes the title 'Aloha', a synopsis, release year (2015), runtime (105 min), cast (Bradley Cooper, Emma Stone, Rachel McAdams, Bill Murray, John Krasinski), director (Cameron Crowe), genres (Drama, Comedy, Romance), and languages (English, Русский). Below the movie card, there are three dropdown menus for user input: 'How familiar are you with this movie?', 'How interested are you in watching this movie via a streaming service?', and 'For unseen movies only: If you were to watch this movie, what would you predict your rating of it to be?'. A blue 'Submit' button is located at the bottom right of the form.

Figure 3.2: Sample Movie Judgement Page from Study Interface

Following these judgments, participants were presented with the movies they chose as having most interest in watching (second question). We named this the ranking task. To select which movies to show to each participant, we started from the list of movies they chose has being “Extremely interested” in watching. If the length of this list was three or more, we showed all such movies. If the count was less, we moved also added the all the movies in the next interest level, which was “Very interested”. This process was continued until the cumulative list contained at least three unique movies.

The ranking task was aimed at finding each participant’s top three items out of these movies in an ordered list. In other words, we asked each participant the following question:

“If you were to pick three movies out of this list, which three would you pick and in what order would you watch them”? We call this ranked list the watch queue. Figure 3.3 demonstrates a sample screenshot of the ranking page from the study interface.

IRMrec participant_2162543

Choose Your Top Picks

	Title	Year	Place in Top Three
	Jumanji	1995	1 2 3
	Heat	1995	1 2 3
	Casino	1995	1 2 3
	Get Shorty	1995	1 2 3
	Leaving Las Vegas	1995	1 2 3

Watch Queue

Rank 1
- 

Rank 2
- 

Rank 3
- 

Submit Queue

Figure 3.3: Sample Movie Ranking Page from Study Interface

For the last part of the first phase, participants were asked to fill out a post-study questionnaire to gather feedback on their overall experience.

3.4.8 Consistency Check

Once each participant completed phase one, they were asked to send us an email and let us know. We also asked them to mention whether or not they are willing to take part in the second phase of the study. As discussed in subsection 3.4.5, for both phases of the study, we included a small set of previously rated movies in the judgment pool for each participant.

This action had two purposes. First, it provided a way to identify participants who were inconsistent in their judgments throughout the study. Second, it offered an additional data point for analysis.

For measurement, Mean Absolute Error (MAE) was utilized. It calculates the average observed error regardless of the value being positive or negative. The error was the difference between a participant rating for a movie during the user study and the rating value for the same movie in their profile. The formula for MAE is:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |r_i - \hat{r}_i|$$

in which r_i shows the existing rating from profile, $\hat{r}_i$ depicts rating captured during the user study, and n is the number of consistency items. We followed below rules to assess eligibility for phase two:

- **MAE $\leq$ 1.0:** Participants who had an MAE of 1.0 or lower were automatically eligible to proceed to phase 2 of the study.
- **MAE $>$ 1.0 and $\leq$ 2.0:** Participants with an MAE greater than 1.0 but less than or equal to 2.0 had their results manually reviewed by us to make a decision.
- **MAE $>$ 2.0:** Participants with MAE greater than 2.0 were automatically rejected from moving to phase 2 of the study since their rating discrepancy was too high.

For all three batches of participants combined, the results of our consistency check performed after the first phase showed a high level of consistency among our participants. The distribution of observed MAE values had an average of 0.35, a median value of 0.3, and a maximum of 0.9.

3.4.9 Phase Two

The second phase was the same as the first one, except that there were approximately 500-700 movies to be judged for each participant. We estimated that it would take approximately 5 hours on average for them to complete this phase. Upon providing judgments for all recommended items, participants were encountered with the ranking page. The difference here was that we showed them all movies they highlighted as having the most interest in watching on this page, regardless of the item belonging to the first or second phase. The last part of the second phase included a post-study questionnaire like before.

3.5 Test Collection Data Cleaning

After the close of data collection, we deleted certain participants' data. First, we removed all participants that did not fully complete either phase 1 or phase 2 of the study. This thesis only examines participants that successfully completed phase 2. Secondly, we mistakenly allowed people with whom I had a personal relationship to participate in the study. We reported this ethics protocol deviation to the university's Office of Research Ethics and received approval to remedy the issue by removing the participants' data from our study. To eliminate these participants, and others that may have learned of the study and then joined MovieLens and our study to opportunistically receive remuneration, we decided that we would exclude participants for which we did not find a pre-existing profile in the ML-32M dataset with a cosine similarity greater than 0.85. As such, we only use participants that we believe were already members of MovieLens prior to the study being announced. The ML-32M dataset was created on October 12, 2023, and our banner announcing the study was posted to the movielens.org site no earlier than Oct 16, 2023.

3.5.1 Study Funnel

We then report on the number of participants at each step. To provide a clear overview of the participant flow through our study, we present the stages of the study process and the number of participants at each stage. These stages show the progress from recruitment to study completion:

1. MovieLens users saw our banner on MovieLens.org.
2. By clicking on the banner, they were redirected to our website.
3. Users could sign up to the study (360 did).
4. Screening was conducted to select eligible participants (271 passed).
5. Consent forms were sent to those who passed screening, and 130 gave consent.
6. From those who gave consent, we asked them to extract and send their MovieLens ratings profiles (113 provided profiles).
7. After producing recommendations, login details were sent to those who sent profiles.
8. The first page participants encountered was the demographics page (107 submitted their information).
9. Participants then completed the knowledge check (quiz) section (103 passed).
10. Afterwards, it was the movie judgments of the first phase (97 successfully completed).
11. Of those who completed phase one, 77 expressed interest in continuing to phase two.
12. Next, 57 participants completed the tasks in the second phase and finished the study.

13. Out of the 57 participants, 6 profiles were removed because we did not find a matching profile for them in ML-32M, and we were left with 51 profiles.

Figure 3.4 is a funnel plot that shows the progression of participants through the study.

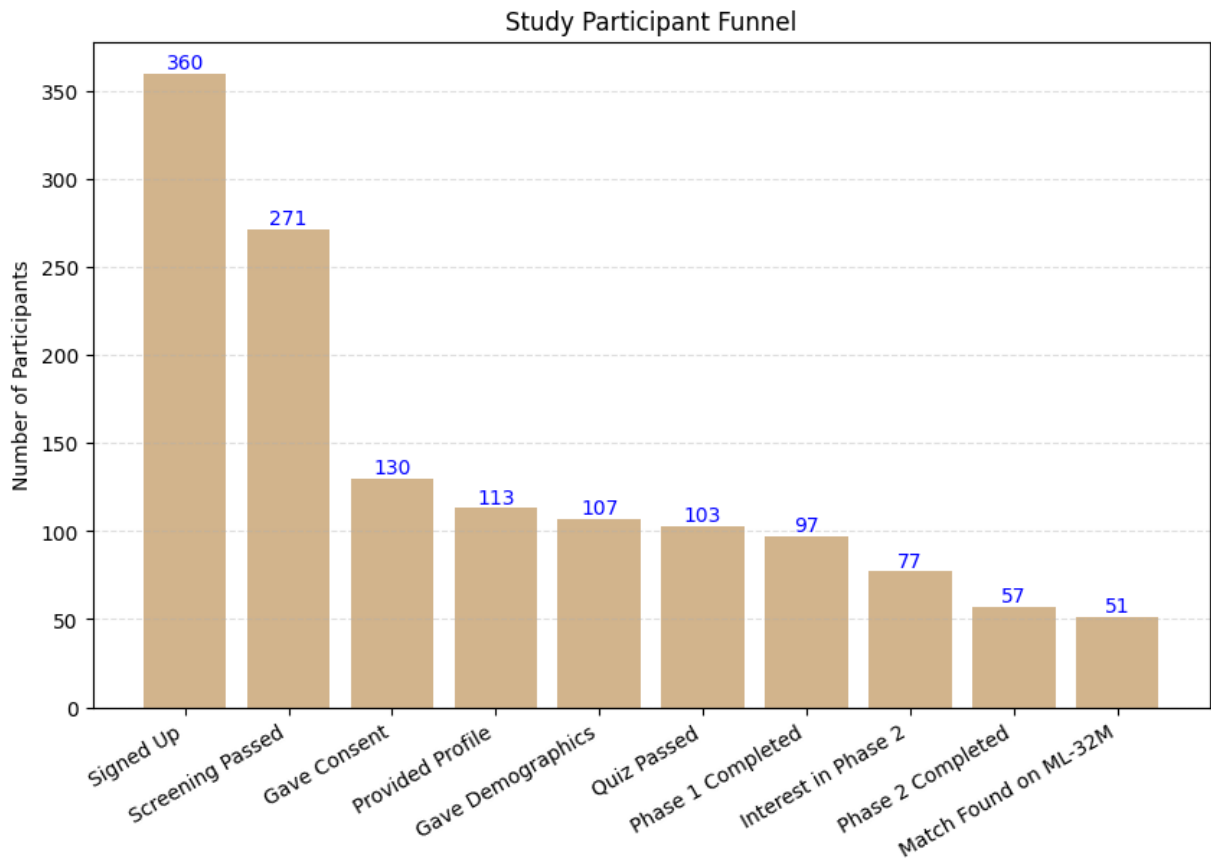


Figure 3.4: Funnel Plot of Study Stages

With the data gathered from our participants during the user study, we continued by performing the data cleaning stage.

3.5.2 Data Preparation

For the first part of analysis (user nature), two different sets of data were prepared to be compared against each other. The first one belonged to the participants of our study. It was decided to only keep those profiles who satisfied two conditions, which were as follows:

- The participants who successfully completed both phases of the user study.
- Participants who had a match probability greater than 85% with their most similar profile in the existing MovieLens dataset, as described in subsection [3.1.2](#).

By considering these filters, we were left with 51 participant profiles. From now on, we designate this subset of participants as selected participants (**SP**). Their judgments from the user study were already available. To gain a better understanding of their differences from the average MovieLens users, we also extracted their historical ratings from the ML-32M dataset. This historical ratings dataset was obtained by filtering ML-32M to only keep those profiles who met the matching threshold with one of our 51 selected participants.

As for the second set of analysis data, and with the goal of having a baseline and representation of the average MovieLens user, we gathered a sample of 10,000 user profiles from the ML-32M dataset alongside their ratings. These profiles were carefully selected so that they did not have any overlap with the other profiles previously identified as possibly belonging to our user study participants. The criteria for selecting these baseline profiles were as follows:

- First, selected those profiles who, after applying the filtering step mentioned in subsection [3.1.2](#) still had more than 20 implicit ratings.
- From all of these eligible users, we randomly selected 10,000 profiles to create the baseline dataset to represent the average MovieLens user.

Chapter 4

Participant Judgment Analysis

4.1 Participant Demographics

This section provides a brief overview of the demographics of our participants. There are four distinct categories to report such as age distribution, gender, ethnicity, and highest level of education. Furthermore, as the ML-1M dataset is the only MovieLens dataset that has demographic information including gender, age, occupation, and zip code, we also report comparisons and analyses for their age and gender.

4.1.1 Age Statistics

We begin by discussing the statistics of participant’s age and their differences with ML-1M users. Table [4.1](#) shows the count, mean, standard deviation, and other relevant metrics for SP and ML-1M.

To complement these, figure [4.1](#) illustrates the age distribution of participants in this study. Please note that per our ethics application, we only recruited adults, and therefore,

Table 4.1: Age Statistics for Selected Participants and ML-1M

Metric	SP	ML-1M
Count	51	6040
Mean Age (Years)	36.86	30.64
Standard Deviation (std)	12.34	12.90
Min	18.00	1.00
25% Percentile	28.50	25.00
Median (50% Percentile)	35.00	25.00
75% Percentile	43.50	35.00
Max	72.00	56.00

there are no minors in our sample. As shown, the majority of participants were in their late 20s to early 40s.

4.1.2 Gender Statistics

We continue by giving a summary of the gender distribution for SP and ML-1M. Table 4.2 demonstrates the frequency percentage of each gender category in both groups.

Table 4.2: Gender Statistics for Selected Participants and ML-1M

Gender	Proportion (SP)	Proportion (ML-1M)
Man	80.4	71.7
Woman	15.7	28.3
Non-binary	3.9	-
Declined to state	-	-

The majority of participants (approximately 80% of SP) identified as men. On the

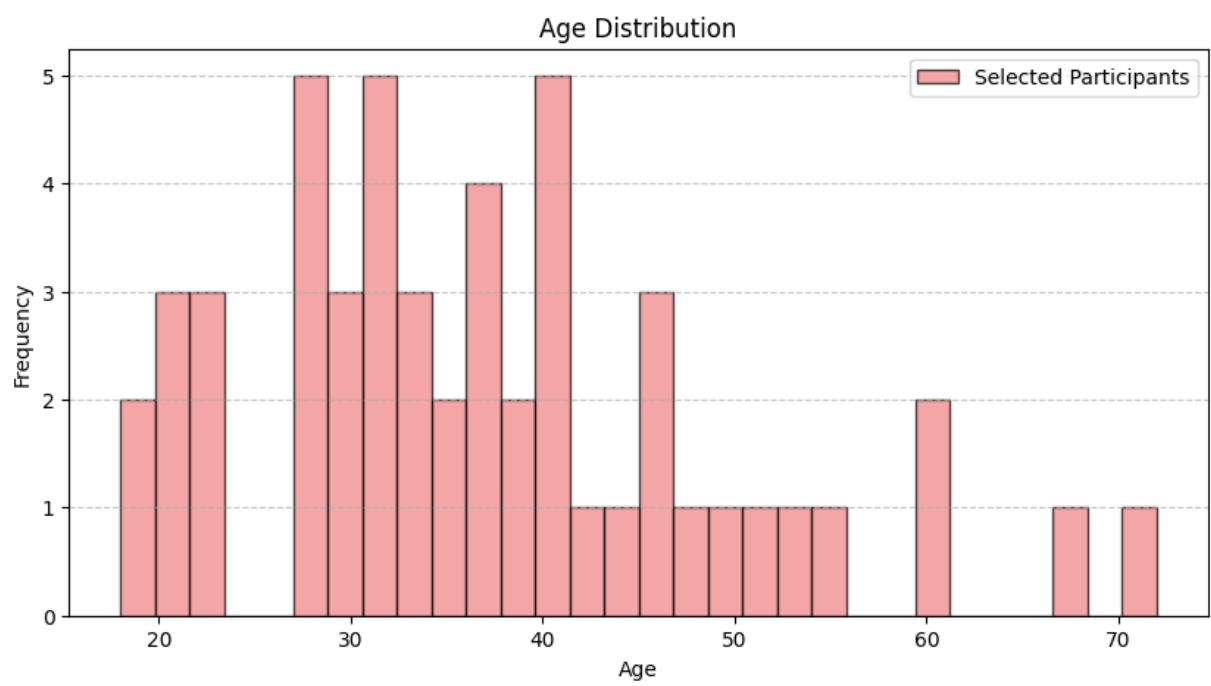


Figure 4.1: Distribution of Participant Age

other hand, women represented about 16% of SP. About 3% of participants was made up of non-binary participants and lastly. As for the ML-1M, it consisted of more users who identified as women (more than 28%) compared to SP.

4.1.3 Ethnicity Statistics

We report on the distribution of ethnicity of our participants, which they self-reported. To provide context, we also include comparable data from Canada and the United States. Table 4.3 shows the proportion of each category of ethnicity.

Table 4.3: Ethnicity Statistics for Selected Participants

Ethnicity	Proportion (SP)
European	86.3%
African	-
Southeast Asian	2%
South Asian	2%
Latin American	4%
Arab/Persian	-
Native American	2%
Declined to state	-
East Asian	2%
Metis/Mixed Ethnicity	2%

The largest subset belongs to participants with European descent (86% in the SP). Those with African descent were not part of the SP. There were also smaller groups of participants who identified as Southeast Asian, South Asian, Latin American, Arab/Persian, Native American, East Asian, and Metis/Mixed ethnicity.

Our pool of participants is less diverse compared to the general populations of the United States and Canada. According to the 2020 U.S. Census ([U.S. Census Bureau, 2021](#)), those identifying as White alone constituted 61.6% of the population. Similarly, the 2021 Canadian Census ([Statistics Canada, 2022](#)) reported that 69.8% of the population identified as White. It’s important to note that the ethnic background information for the MovieLens users (ML-32M) is not available. This makes it challenging to know whether our participant pool is an accurate reflection of the typical MovieLens user demographic.

4.1.4 Education Statistics

We conclude this section by providing a summary of educational background of our participants. Table 4.4 presents the proportion of participants in each educational category.

Table 4.4: Education Statistics for Selected Participants

Education Level	Proportion (SP)
High School	10%
Some College	25.5%
Bachelor’s	35.3%
Master’s	19.6%
Postgraduate	3.9%
Doctorate	5.9%

The results show that the majority of participants hold a Bachelor’s degree. After that, the next most common educational levels are Master’s degrees and some college experience. There is a relatively smaller pool of participants with high school education, postgraduate degrees, and doctorates.

Compared to 2022 U.S. Census data ([U.S. Census Bureau, 2021](#)) and 2021 Canadian

Census ([Statistics Canada, 2022](#)), our participant pool has a higher level of educational attainment compared to the general populations of the United States and Canada. The proportion of people with a Bachelor’s degree or higher was around 33% in Canada and 37% in the U.S., whereas in our study, this proportion ranges from 63.7% to 64.7% when combining Bachelor’s, Master’s, Postgraduate, and Doctorate categories. Therefore, our participants were significantly more likely to hold bachelor’s degrees and advanced degrees compared to the general population.

4.2 Nature of Movielens Users & Participants

The next group of analyses is based on the two datasets from section 3.5.2. The first dataset consists of the historical ratings of selected participants. This means that we compiled all previous ratings of our selected participants from the ML-32M dataset. The second dataset belongs to the ratings from the sample of 10,000 users of the ML-32M dataset, which is aimed to represent the average Movielens user. In this section, our focus is to compare various characteristics of these two groups and uncover their intrinsic differences.

4.2.1 Number of Ratings Per User

The first part of comparing the nature of users for the two datasets is to compare the number of ratings provided by users and identify any differences between these two groups. With this goal in mind, table 4.5 showcases the statistics.

It can be observed from the data that our selected participants tend to have significantly more ratings than the average MovieLens user. Figure 4.2 demonstrates the distribution of ratings for both groups. To sum up, the findings suggest that our selected participants are generally much more engaged with the Movielens website compared to a typical user.

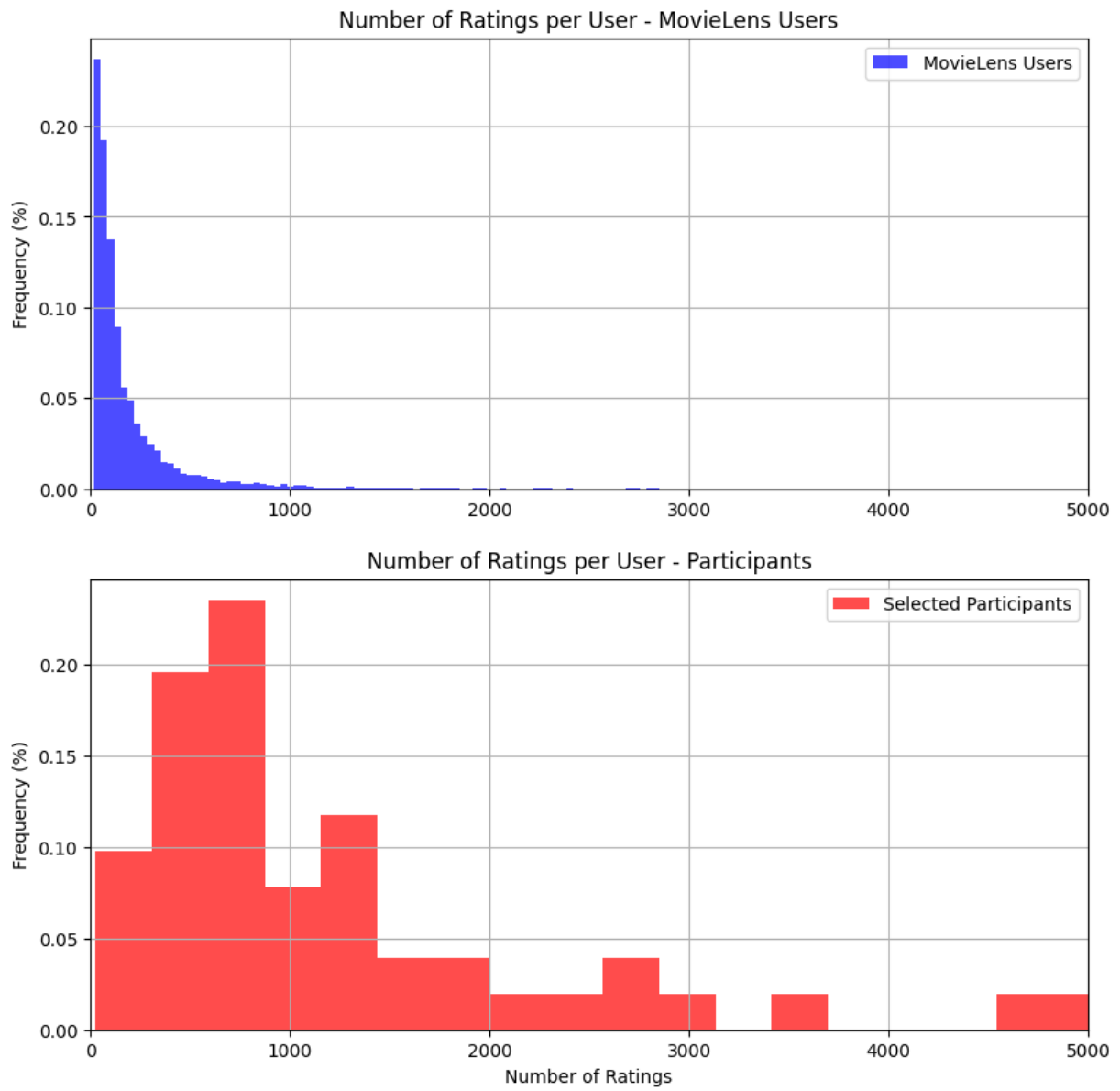


Figure 4.2: Distribution of Number of Ratings per Person -
Movielens Users (top) and Selected Participants (bottom)

Table 4.5: Number of Ratings Per User for MovieLens Users and Selected Participants

Metric	MovieLens Users (10K)	Selected Participants (SP)
Count	10,000	51
Mean	190.9	1425.1
Standard Deviation (std)	276.1	1567
Min	20	28
25% Percentile	55	561.5
Median (50% Percentile)	102	849
75% Percentile	211	1552
Max	5017	8496

4.2.2 Distribution of Rating Values

Next, we looked into analyzing the distribution of rating values for both datasets that were elaborated at the top of this section (4.2), meaning that for selected participants, we used their rating profiles as provided to us. Table 4.6 shows the percentage values of ratings for each group.

For this part, macro averaging was utilized instead of micro averaging. In this approach, the distribution of rating values is calculated for each participant and then, the average of all these distributions is considered for the output. One of the advantages of this method is that it helps reduce the skewing effect of participants who have very high or low number of ratings. This way, the results are compiled in a way to account for the variations in rating behavior between different users. The results of average rating distribution per user for both groups are shown in figure 4.3. There are a few key differences between MovieLens users and selected participants in terms of their distribution of rating values:

1. **Low Rating Values:** Our participants have a higher historical frequency of very

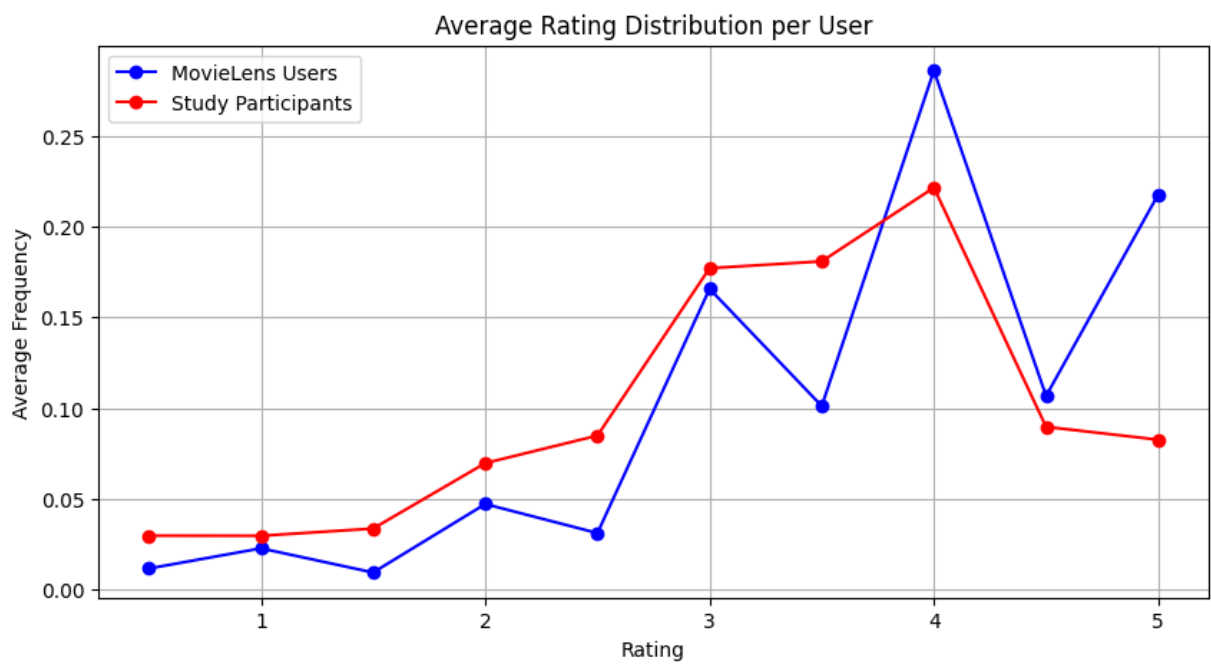


Figure 4.3: Average Rating Distribution per User

Table 4.6: Distribution of Rating Values for MovieLens Users and Selected Participants

Rating Value	MovieLens Users (%)	Selected Participants (%)
0.5	1.6%	4.8%
1.0	2.7%	3.9%
1.5	1.6%	4%
2.0	6%	8.6%
2.5	5.2%	9.6%
3.0	18.5%	17.8%
3.5	13.7%	16.7%
4.0	26.7%	18.3%
4.5	9.5%	7.4%
5.0	14.4%	8.7%

low ratings, such as 0.5, 1.0, 1.5 in comparison to MovieLens users.

2. **Middle and High Ratings:** Both groups show a high number of middle to high ratings (3.0, 3.5, 4.0). However, the trend for selected participants is more consistent, whereas MovieLens users have a peak at 4.0.

3. **Perfect Ratings:** In general, MovieLens users are more likely to give perfect scores to movies compared to selected participants, as the average frequency of 5.0 for the MovieLens users is more than double of that for participants.

4. **Use of Half Ratings:** It is evident that participants tend to make more careful use of the half ratings (3.5, 2.5, etc.). On the other hand, MovieLens users are more likely to have integer value ratings. However, there's a possibility that some MovieLens users in our dataset come from a period when the platform only allowed integer ratings.

4.2.3 Distribution of Average Ratings

Furthermore, we investigated the distribution of average rating per user for each one of the two groups described in the beginning of Section 4.2. Figure 4.4 shows the results for the both Movielens users and selected participants. The two distributions of this part reinforce the previous finding that participants are generally more inclined towards providing lower ratings.

4.2.4 Last Year of Movie Rated

For the next piece of analysis, we made a distribution for the last year of movie rated of both groups (4.2). For each person, this is the largest value of the release year of any movie in their rating profile. Figure 4.5 shows the results for the Movielens users. There are two peaks for the last 1990s and mid 2015. As for the participants, 2020 was the earliest date with last seen values of participant rating. In other words, all the participants provided a rating in 2020 or after. As presented by table 4.7, more than 85% of selected participants were active in 2023.

Table 4.7: Distribution and Count of Last Year of Movie Rated for Selected Participants

Year	Count	Percentage
2020	1	2%
2022	6	11.8%
2023	44	86.3%

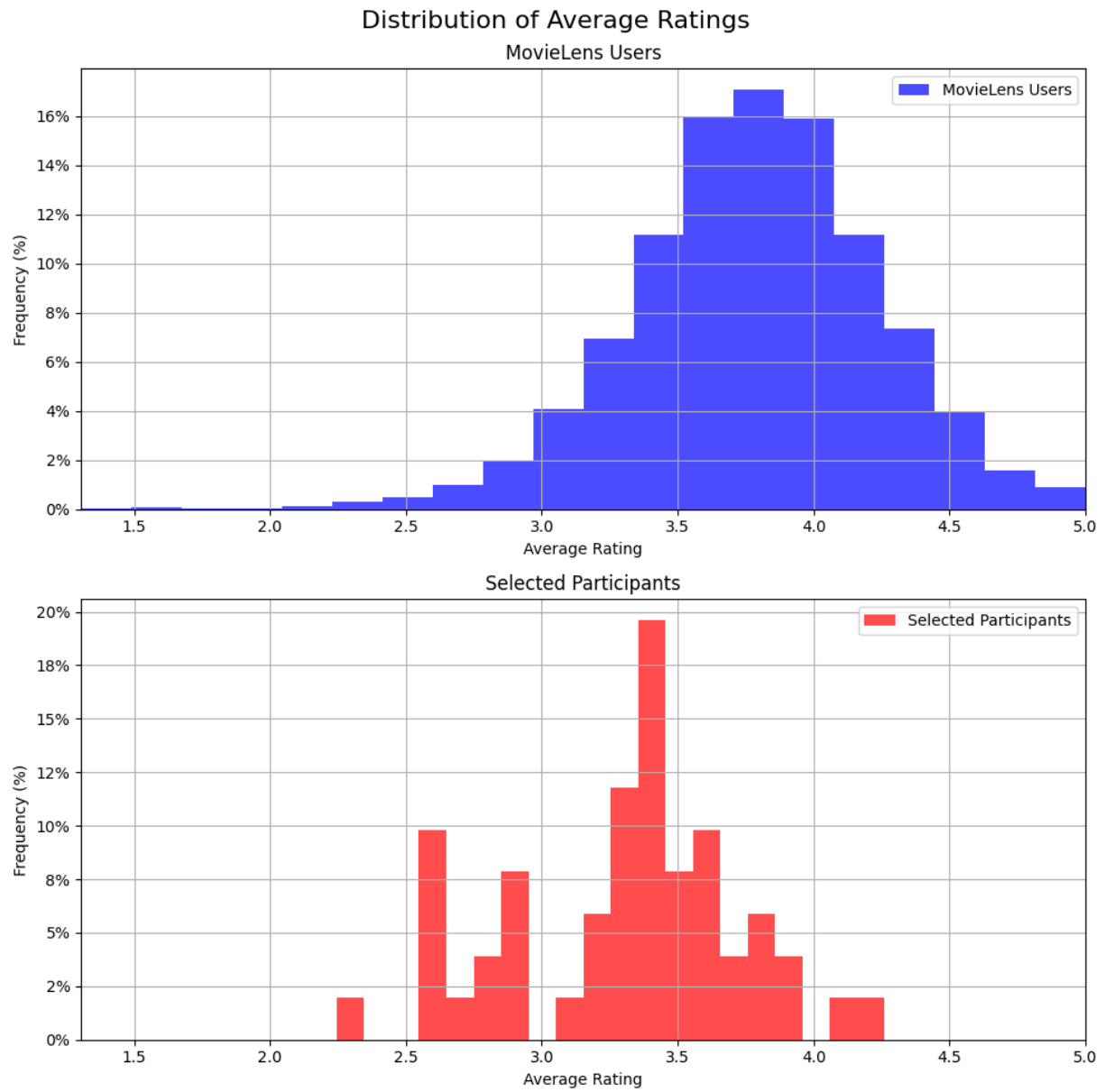


Figure 4.4: Distribution of of Average Rating Distribution -
Movielens Users (top) and Selected Participants (bottom)

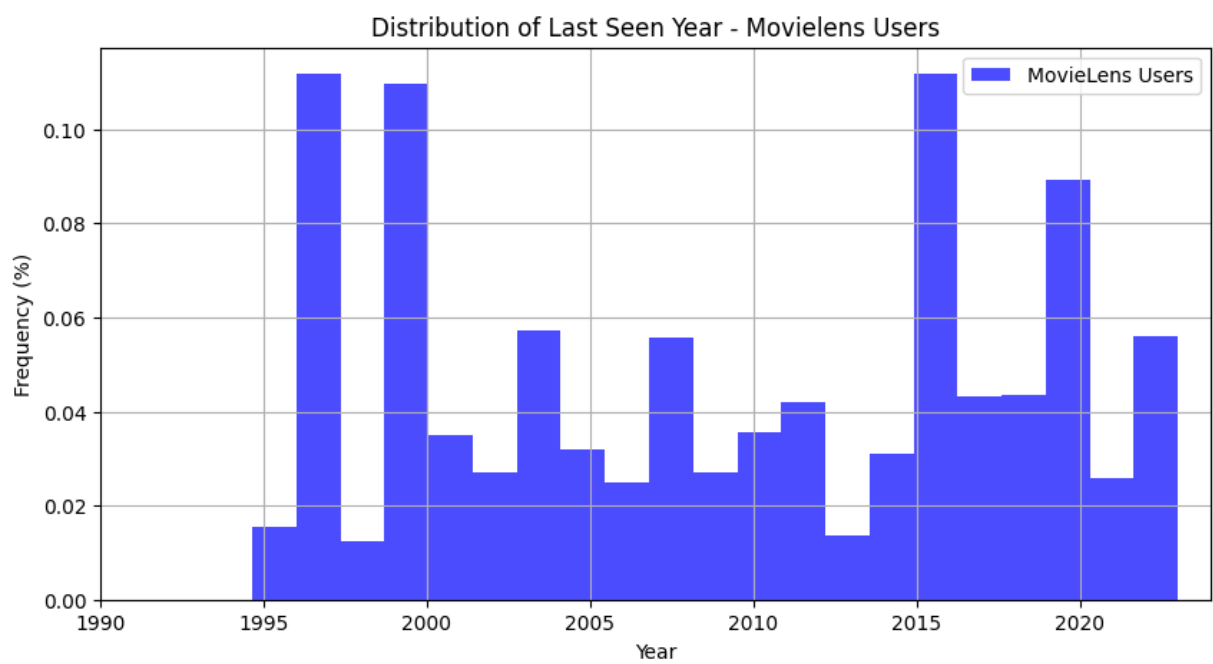


Figure 4.5: Distribution of Last Year of Movie Rated - Movielens Users

4.2.5 Year Distribution of Rated Movies

Second, we analyzed the distribution of the release years of movies rated by the selected participants and the MovieLens users. We followed the macro-averaging method like before, meaning that we calculated the average year for each individual, and then took the average of these averages. Table 4.8 summarizes the average and median year of the rated movies.

Table 4.8: Year Statistics of Rated Movies for MovieLens Users and Selected Participants

Metric	MovieLens Users (10K)	Selected Participants (SP)
Average Year	1995.7	2000.9
Median Year	1995.1	2002

Furthermore, figure 4.6 illustrates the distribution of the average movie years for both groups.

It can be seen from the data that the median year of rated movies for Movielens users is 1995, while for the selected participants, it is 2002.

4.2.6 User Lifespan

For the last part of this section, we analyzed the difference between the lifespans of the two datasets mentioned at the start of section 4.2. We define the lifespan as the time between a user’s first rating and their last rating. The rating profiles of both sample Movielens users and selected participants were available from Chapter 3. This means that for our participants, we estimate their lifespans based on their matching profile we found in the ML-32M dataset. Table 4.9 presents the summary statistics for both distributions.

Figure 4.7 shows the distribution of lifespans for both the Movielens sample users and the study participants. The results show a significant difference in engagement between

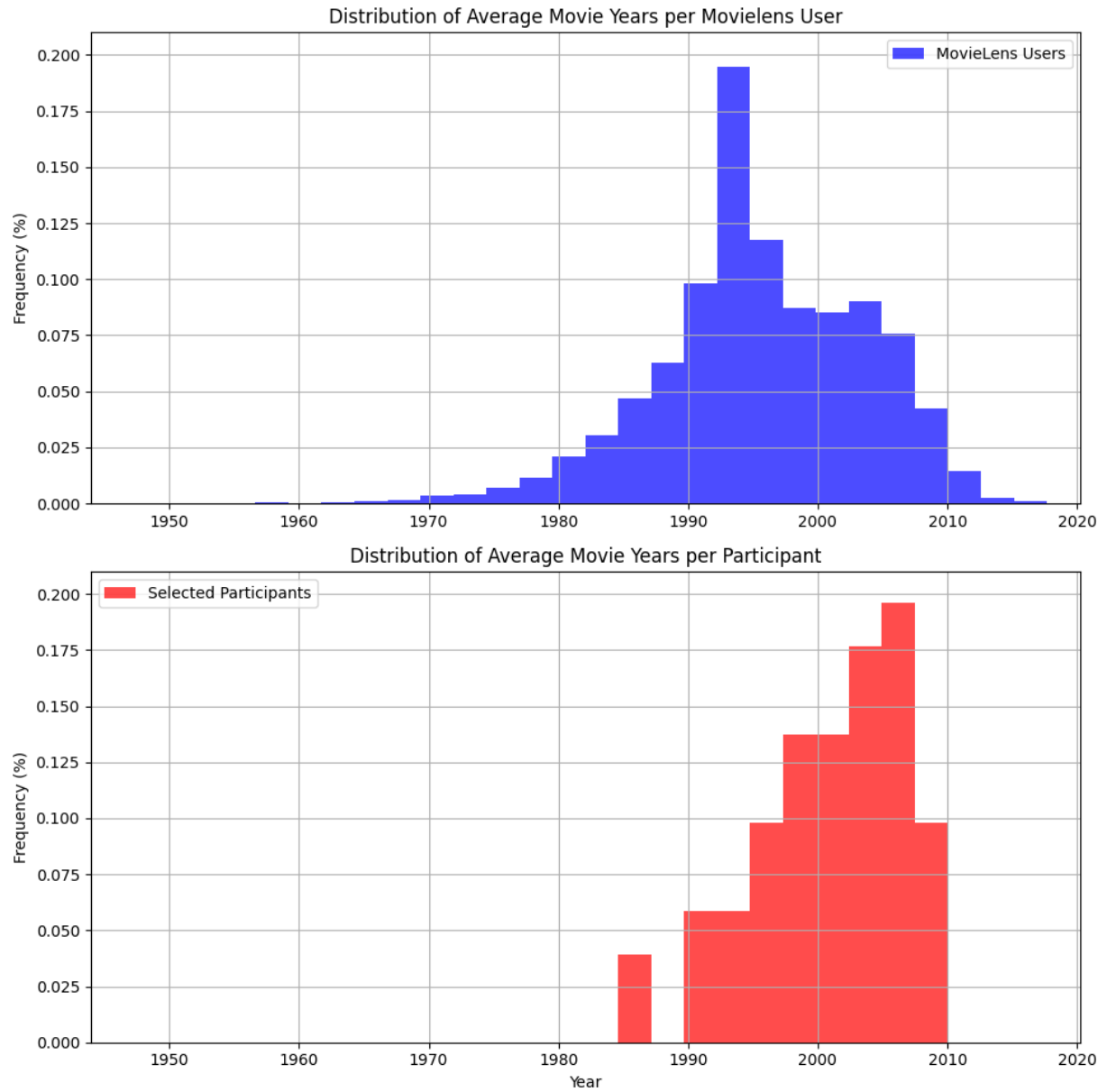


Figure 4.6: Distribution of Release Year of Rated Movies -
Movelens Users (top) and Selected Participants (bottom)

Table 4.9: Summary of User Lifespan

Metric	Selected Participants	Movielens Users
Count	51	10,000
Mean	7.8 years	10.5 months
Standard Deviation	7.2 years	2.5 years
Minimum	20 days	0
25th Percentile	2.2 years	27.7 minutes
Median	5.2 years	15.8 hours
75th Percentile	11.9 years	17.7 weeks
Maximum	23.8 years	26 years

the participants and the general Movielens user base. Firstly, participants have a much longer average lifespan (about 7.8 years on average) compared to the general Movielens user base of which 75% are only around less than 5 months. Second, median values show half of the Movielens user base don't use the system after their first day. These findings demonstrate that our selected participants form a very engaged subset of Movielens users who have a prolonged and sustainable level of engagement with Movielens.

4.2.7 Implications

The comparison between our study participants and random MovieLens users shows important implications for recommender system evaluation. Our findings reveal that these two groups are different, and these differences need to be taken into account when evaluating a system. Our participants appear to mostly represent movie enthusiasts. This distinction is evident in several aspects. For instance, the median number of ratings for random users was 102, whereas the median for participants was 849. Furthermore, we showed that

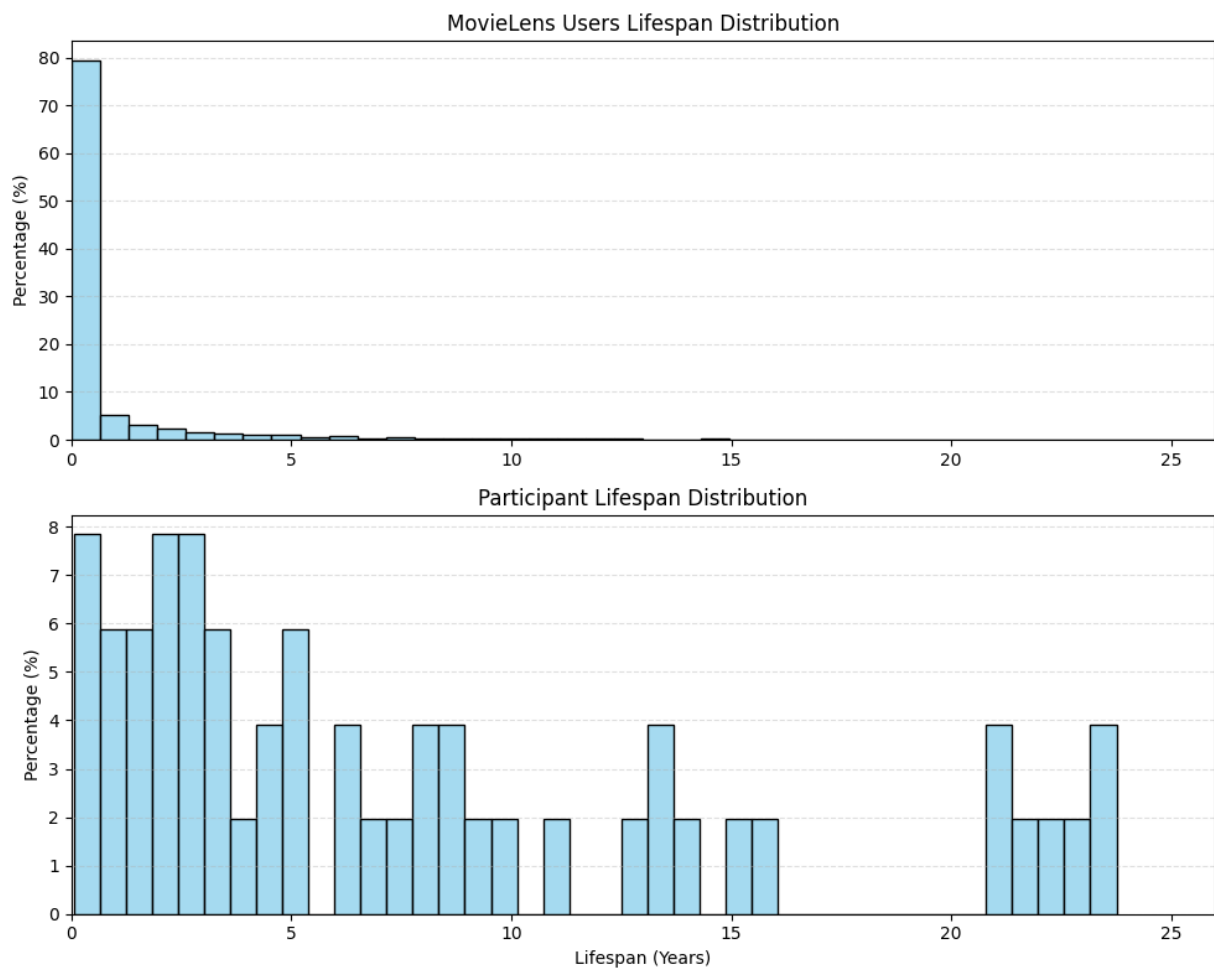


Figure 4.7: Distribution of Lifespans -
Movielens Users (top) and Selected Participants (bottom)

participants had a much longer lifespan (median of around 1,900 days vs. only 1 day for random users). Also, participants made more use of half-star ratings and were less inclined towards giving perfect ratings. Therefore, we suspect that our participants are a better representation of regular, engaged MovieLens users.

Our test collection is likely representative of active MovieLens users, while traditional test/train methods would also represent the large number of users who enter the system, get recommendations, and never return. This needs to be considered when evaluating a recommender system. Our test collection may not be helpful if it is important to evaluate the recommendation ability to those who may not become regular users.

Furthermore, another important factor to consider is time. Our test collection represents current users, while a majority of the MovieLens ratings is from users who haven't been active for many years. This raises questions about the relevance of evaluating recommendation performance for users who were last active years ago. For instance, measuring the ability to recommend movies to someone last seen in 2001 may not be meaningful for current model performance.

4.3 Consistency Checks

Furthermore, we looked into the rating behavior of our participants and, more importantly, how consistent they were when providing ratings for the items that were part of their rating profiles. Two sets of analyses—test-retest reliability and proportion of consistency movies tagged as seen before by participants—were performed with respect to this section, which are referred to as consistency checks. These checks are essential for better understanding the reliability of user study judgments.

4.3.1 Test-Retest Reliability

As detailed by [Amatriain et al. \(2009\)](#), test-retest reliability is a measure to find out how consistent participants are in their behavior, which is translated to rating, over multiple sessions. This consistency is studied by using the Pearson correlation coefficient. For this study, we defined the two different sessions as the ratings given during the study and the historical ratings of the same users from the MovieLens dataset. The macro-averaged correlation coefficient was 0.81. This indicates a high consistency in participant ratings over the two sessions. Table 4.10 presents the detailed statistics of the correlation coefficients across participants.

Table 4.10: Test-Retest Reliability Statistics

Metric	Value
Mean	0.808
Standard Deviation	0.143
Minimum	0.446
25th Percentile	0.716
Median (50th Percentile)	0.839
75th Percentile	0.919
Maximum	1.000

The minimum correlation of 0.446 suggests that even the least consistent participant showed a moderate level of reliability. The highest correlation of 1 shows perfect consistency for at least one of the participants. [Amatriain et al. \(2009\)](#) also reported that their overall test-retest reliability was 0.924, which they considered high as any value over 0.9 is usually considered “good” in classical test theory. Our average correlation was 0.808. Although showing a strong consistency, it was lower than their overall reliability. However, it is

important to note that their study has a different design to ours as they had three trials over time and calculated both reliability and stability.

Our results show that most participants maintained considerable consistency in their ratings. Some showed very high consistency (correlations above 0.9) and others had lower (but still positive) consistency.

4.3.2 Rate of Past Views for Consistency Items

This part examines how often participants identified the consistency items as movies they had previously seen. After performing a macro-averaging among participants, we found that the mean proportion of consistency items tagged as seen was 0.88. Table 4.11 presents the detailed statistics of this relationship across participants.

It's important to note that a rating on MovieLens doesn't necessarily mean a user has previously watched the movie. Some users may rate movies based on trailers, reviews, or other factors without having seen the film. The high correspondence value of 0.88 suggests that in most cases, ratings do align with viewing history, but it's not a perfect indicator.

The distribution of values shows variation among participants. The median rate of 0.94 indicates that half of the participants reported having seen at least 94% of their consistency check movies. However, rates ranged from a minimum of 0.34 to a maximum of 1.00. The variations could be the result of a few factors such as the amount of time since they first provided the original rating (which leads to forgetting), confusing similar movie titles with each other, and the fact that our assumption is not accurate and participants have rated movies on MovieLens that they hadn't actually watched.

Table 4.11: Statistics of Consistency Item Past Views

Metric	Value
Number of Participants	51
Mean Recognition Rate	0.878
Standard Deviation	0.160
Minimum	0.340
25th Percentile	0.860
Median (50th Percentile)	0.940
75th Percentile	0.980
Maximum	1.000

4.4 Participant Feedback Analysis

Another piece of analysis carried out was to take a closer look at the feedback of our participants and categorize them into common themes. This will greatly help future related research to improve upon this study and also find interesting research questions to look into. Based on all gathered feedback, which is available in appendix B, here is a summary of topics that were frequently observed:

- **Enjoyment and Positive Experience:** Six participants expressed their “satisfaction” or found the study “interesting”. Furthermore, three participants found the process helpful for finding out about new movies to watch.
- **User Interface Issues:** Three participants had issues with the “submit button becoming invisible”, which turned out to be caused by their high browser zoom level. Also, some suggested adding a “progress bar” to the list of rated items page. Lastly, there were two requests for a back button or the ability to edit a judgment right after submitting it.

- **Content-Related Feedback:** Five participants said that recommendations included “non-traditional movies” like short films, TV shows, and comedies, which they found confusing. Three suggested the ability to exclude certain categories or genres from the recommendations.
- **Rating System and Movie Judgments:** Three participants found it difficult or “non-intuitive” to rate unseen movies solely based on provided information. As a workaround, some suggested including more information or trailers to help in judgment.
- **Language and Cultural Aspects:** Two participants mentioned issues with foreign language films. For instance, participant mentioned that they normally “disqualify foreign movies from their watch list”.
- **General Suggestions for Improvement:** Three participants offered specific suggestions for improving the study or interface such as: including “movie synopses” in the final ranking list to help remember the movies, avoiding “duplicate movies from the same series or remakes”, providing context on how participants “capture and consume media”. Specifically, a participant suggested this information influences their ranking of movies to watch next.
- **Interest in Data and Results:** A few participants expressed interest in seeing the results or insights from the study.

4.5 Pooling Depth & Number of Judgements

An important part of our process was the pooling, which set the number of recommendations each participant would judge. This section provides an overview of the number of

judgments made by participants and the depth of pooling needed to generate these recommendations. We considered the SP dataset for this section and started by analyzing the number of movie judgments provided. The statistics are as shown in table 4.12.

Table 4.12: Statistics of Number of Judgements and Pooling Depths

Metric	Number of Judgements	p1 depth	p2 depth
Number of Participants	51	51	51
Mean	670.4	11.9	63.2
Standard Deviation	24.2	2.1	12.5
Minimum	634	10	50
25th Percentile	661.5	10	53
Median (50th Percentile)	664	11	60
75th Percentile	665.5	13	71
Maximum	775	19	94

It is seen that most participants judged between 661 and 666 movies and the median was 664. The relatively small standard deviation (24.20) suggests consistency in the number of judgments for participants.

In addition to the number of judgments, we also analyzed the pooling depths for both the first phase (p1) and the second (p2). The p1 depths had a mean of 11.9 with a standard deviation of 2.1, while the p2 depths had a mean of 63.2 with a standard deviation of 12.5. For p1, most depths fell between 10 and 13, with a median of 11. For p2, the range was much wider, with most depths falling between 53 and 71, and a median of 60.

4.6 Data & Participant Insights

In this section, we are delving deeper into the many interesting insights that can be inferred from the judgments of our participants throughout the user study. For all the analyses in this section, we are using the SP dataset as detailed in section 3.5.2, which consists of the data of 51 participants. Before proceeding with our analysis, it’s essential to introduce a part of the data collected. After excluding the consistency check movies from the SP dataset, the majority of movies (88.8%) were reported as unseen by our participants and only 11.2% were reported as seen. This distribution of seen versus unseen movies provides important context for our subsequent analyses. By examining aspects such as the desire to watch for multiple categories, the effect of perceived ratings, and the role of familiarity, we aim to provide a better understanding of participant behavior and preferences.

4.6.1 Interest for Unseen Movies

Our first research question explores the relationship between participants’ interest levels and their predicted ratings for movies they have not seen before and are not part of the consistency check. To analyze this, we first calculated the number of occurrences for each pair of rating and interest level. Then, we normalized these numbers by dividing them by their total count for each predicted rating value. In other words, we performed micro-analysis of the data so that each movie rating contributes equally to the final outcome. The results of this analysis are demonstrated in table 4.13.

Several points of interest can be observed from the results. As predicted ratings increase, we can see a change in the distribution of interest levels. For predicted ratings of 0.5 to 2.0, the dominant interest level is “Not interested” and it accounts for over 80% (marginal) of responses. As rating values increase, this dominance fades away. For mid-range predicted ratings (2.5 to 3.5 stars), we observe a more diverse distribution of interest levels.

Table 4.13: Interest Levels vs. Predicted Rating for Unseen Movies (%)

Rating Value	Not interested	Somewhat interested	Interested	Very interested	Extremely interested	Total
0.5	3.9	0.0	0.0	0.0	0.0	3.9
1.0	3.5	0.2	0.1	0.0	0.0	3.8
1.5	4.4	0.6	0.1	0.0	0.0	5.1
2.0	8.4	1.6	0.2	0.0	0.0	10.3
2.5	8.7	5.9	1.0	0.1	0.0	15.8
3.0	6.4	11.5	4.9	0.8	0.1	23.7
3.5	1.6	5.1	8.8	3.4	0.4	19.3
4.0	0.4	1.2	3.7	6.2	2.1	13.6
4.5	0.1	0.1	0.4	1.4	1.7	3.7
5.0	0.0	0.0	0.0	0.2	0.7	0.9
Total	37.4	26.1	19.3	12.2	5.0	100.0

“Somewhat interested” and “Interested” categories become more prominent (“Somewhat interested” peaks around the 3-star rating). At higher predicted ratings (4.0 and above), “Very interested” and “Extremely interested” levels become absolutely prevalent.

These patterns suggest that while there is a general trend of increasing interest with higher predicted ratings, the relationship is not linear. Each interest level corresponds to a range of predicted ratings, and these ranges do overlap.

4.6.2 Interest for Seen, Absent from Movielens

It is evident that there are a number of movies that each participant has seen but has not yet provided a rating for on Movielens. This means that there is no previous record of interaction between that participant and the movie. For simplicity, we refer to these as seen-unrated movies. After filtering the data for these movies, we followed the same process as described in subsection 4.6.1. Table 4.15 represents the distribution of predicted rating values for each one of the interest levels.

Table 4.15: Interest Levels vs. Predicted Rating for Seen, Unrated Movies (%)

Rating Value	Not interested	Somewhat interested	Interested	Very interested	Extremely interested	Total
0.5	1.0	0.0	0.0	0.0	0.0	1.0
1.0	1.0	0.0	0.0	0.0	0.0	1.0
1.5	1.4	0.1	0.0	0.0	0.0	1.5
2.0	4.5	0.5	0.0	0.0	0.0	5.0
2.5	7.7	1.5	0.1	0.0	0.0	9.3
3.0	14.2	6.2	4.1	0.8	0.0	25.3
3.5	8.5	9.1	4.4	2.4	0.2	24.6
4.0	6.5	6.1	5.6	2.8	0.7	21.8
4.5	0.9	1.6	1.7	1.7	1.0	6.9
5.0	0.4	0.5	0.9	0.6	1.3	3.7
Total	46.0	25.5	16.9	8.3	3.3	100.0

Comparing these results to those for unseen movies shows some interesting differences. First, for lower rating values (3 and below), “Not interested” is more dominant for seen-

unrated movies than for unseen ones. This suggests that participants are less likely to want to rewatch movies they didn't enjoy. Second, the move from "Not interested" to higher interest levels happens at higher ratings for seen-unrated movies (around 3.5 stars) compared to unseen movies (around 2.5 stars). This indicates a potentially higher threshold for rewatching. Next, for high ratings (4 to 5 stars), the distribution of interest levels is more evenly spread for seen-unrated movies. On the other hand, "Very interested" and "Extremely interested" are more dominant for unseen movies.

4.6.3 Interest for Seen, Consistency Check Movies

We continued by assessing the extent to which participants were interested in watching another category of movies given their ratings for them. This category consisted of those movies that participants tagged as seen and were added by us as part of the consistency check process. From the distribution of predicted ratings per level of interest that is depicted in table 4.17, it can be observed that most participants are not interested in watching a movie in this category if they think they will rate it 3.0 or less. They are mostly somewhat interested in a movie if they predict a rating of 3.5 or 4, and if their anticipated rating is 5.0, they are likely to be extremely interested.

Table 4.17: Interest Levels vs. Predicted Rating for Seen, Consistency Check Movies (%)

Rating Value	Not interested	Somewhat interested	Interested	Very interested	Extremely interested	Total
0.5	1.6	0.0	0.0	0.0	0.0	1.6
1.0	2.1	0.0	0.0	0.0	0.0	2.2
1.5	2.6	0.1	0.0	0.0	0.0	2.7
2.0	5.7	0.5	0.2	0.0	0.0	6.4
2.5	7.1	2.2	0.6	0.2	0.0	10.1
3.0	10.0	5.1	1.7	0.4	0.1	17.3
3.5	6.1	8.6	4.3	1.1	0.1	20.3
4.0	4.0	6.7	5.5	2.9	0.6	19.7
4.5	0.7	2.7	3.9	2.8	1.6	11.7
5.0	0.8	1.3	1.9	2.0	2.1	8.0
Total	40.9	27.2	18.1	9.4	4.5	100.0

4.6.4 Interest for All Seen Movies

A check was also done for all recommendations that were chosen as being seen before, regardless of other aspects. The results are presented by table 4.19. We can observe that the distribution has a similar trend with the seen, consistency check movies from subsection 4.6.3. This indicates that when a movie is seen before by a participant, there is not much difference in interest if it was in the consistency check items or not.

Table 4.19: Interest Levels vs. Predicted Rating for All Seen Movies (%)

Rating Value	Not interested	Somewhat interested	Interested	Very interested	Extremely interested	Total
0.5	1.3	0.0	0.0	0.0	0.0	1.3
1.0	1.5	0.0	0.0	0.0	0.0	1.5
1.5	2.0	0.1	0.0	0.0	0.0	2.1
2.0	5.0	0.5	0.1	0.0	0.0	5.5
2.5	7.3	1.8	0.3	0.1	0.0	9.4
3.0	12.3	5.8	3.0	0.6	0.0	21.8
3.5	7.6	8.7	4.3	1.8	0.2	22.5
4.0	5.4	6.3	5.7	2.9	0.6	21.0
4.5	0.9	2.0	2.7	2.3	1.3	9.2
5.0	0.5	0.8	1.4	1.3	1.7	5.8
Total	43.7	26.0	17.5	9.0	3.8	100.0

4.6.5 Interest for Seen vs Unseen Movies

Furthermore, we analyzed the difference between the distribution of interest levels for seen ratings and unseen predicted ratings. The difference was obtained by calculating the percentage of each interest level and rating pair for both seen and unseen categories and then subtracting the corresponding values. The goal was to check if we consider a rating, does having seen a movie affect one's desire to watch it compared to a movie they haven't seen yet? The results of the analysis are visualized in the heatmap shown in Figure 4.8. Positive values (cells in red) show a higher percentage for seen movies, while negative values (blue cells) indicate a higher percentage for unseen movies.

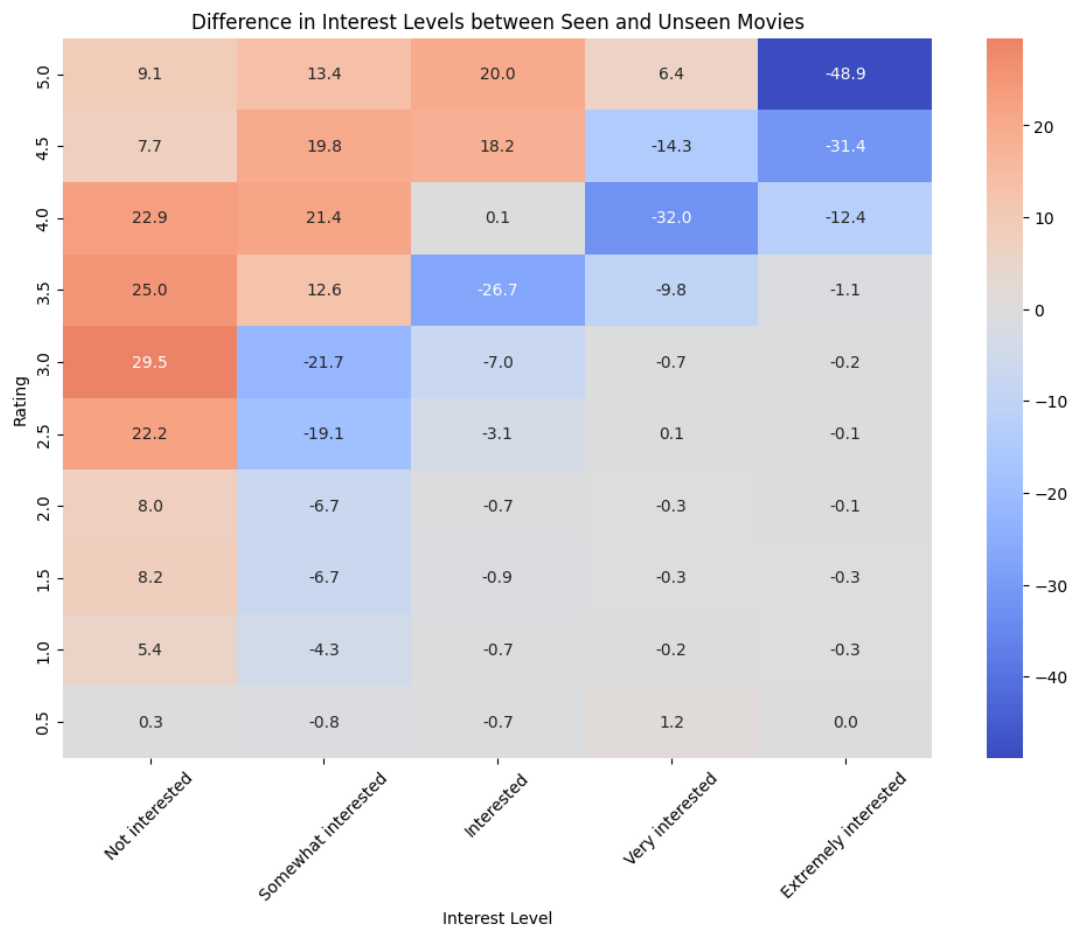


Figure 4.8: Difference in Interest Levels between Seen and Unseen Movies

There are a few conclusions that can be drawn from the results. For highly rated movies (4.5-5.0 stars), participants show less interest in rewatching seen movies, but express strong interest in watching unseen ones. This trend is weaker but still present for movies rated 3.5-4.0 stars.

As for the lower-rated movies (3.0 stars and below), participants are generally less interested in rewatching seen movies compared to trying unseen ones with similar ratings. This points out to a slight openness to new experiences, even with potentially lower-rated content.

Lastly, the “Extremely interested” level maintains a higher percentage for unseen movies among all ratings. This difference is more significant for higher rated movies. This suggests a strong novelty effect in which participants are excited about watching new movies they expect to rate highly.

4.6.6 Interest vs Familiarity

We continued by assessing the distribution of interest levels for each category of familiarity that the participants chose. To ensure the integrity and validity of the results, those movies that were added as part of the consistency check were first removed. Table [4.21](#) presents the distribution.

The data reveals some interesting insights. For the familiarity levels of “Seen” and “Never heard of”, the absolute dominant interest level is “Not interested”. This suggests that, due to a potential lack of novelty, participants are not really interested in re-watching a movie they have seen.

This could show the challenge of recommending interesting but unfamiliar movies. However, it’s important to note there might be other factors in play. For instance, it’s possible that recommendations viewed as less relevant by participants usually include unfamiliar

Table 4.21: Interest Levels vs. Familiarity (%)

Familiarity Level	Not interested	Somewhat interested	Interested	Very interested	Extremely interested	Total
Seen	5.1	2.8	1.9	1.0	0.4	11.2
Very familiar	2.3	2.7	3.1	3.0	2.2	13.3
Familiar	5.7	6.5	6.5	4.1	1.4	24.2
Never Heard of	25.3	13.9	7.6	3.7	0.9	51.4
Total	38.3	26.0	19.1	11.8	4.8	100.0

movies, or that participants are less likely to show interest in movies they don’t know at all. In contrast, items tagged as “Very familiar” or “Familiar” (but not seen) show a more balanced distribution. They have higher percentages in the “Interested”, “Very interested”, and “Extremely interested” categories compared to seen movies. This suggests a positive correlation between familiarity and interest in unseen movies.

4.6.7 Interest with Low Predicted Rating

There were encounters of a situation in which a participant showed a good level of interest in watching a recommendation, but their predicted rating was low. Therefore, a question was formed to check if participants want to watch movies they think they’ll give a low rating to. To provide a full description, we combine examples with the discussion:

- **Historical or Controversial Content:** The most prominent example is one participant’s interest in “Adolf Hitler: The Greatest Story Never Told”. They gave it the lowest predicted rating of 0.5 but were “Very interested” in watching it. According to [Wikipedia \(2024\)](#), this movie is a “revisionist and pro-Adolf Hitler film” promoted

by neo-Nazis. It is reasonable that someone might be interested in seeing such a movie out of curiosity or to be able to critique it themselves, but they would not want to express this interest in watching the movie in terms of a favorable rating or even rate the movie at all.

- **Cult Classics and Infamous Movies:** One participant showed high interest in “The Room” (2003) but gave it a 0.5 rating. This film is known for being “so bad it’s good” that it attracts viewers exactly because of its poor quality. This suggests an interest in seeing culturally significant or notorious movies without considering their perceived quality.
- **Horror and Thriller Genres:** Many recommendations were horror or thriller movies for which participants showed high interest but predicted low ratings. For example: “House of 1000 Corpses” (rating 1.0, Extremely interested), “The Texas Chain Saw Massacre” (rating 1.5, Extremely interested), and “Eden Lake” (rating 1.0, Very interested). This could mean that for some participants the experience of watching horror movies is interesting, even if they don’t expect to give a high rating.

These instances suggest that there are many scenarios where a person does not expect to like a particular movie but is interested in watching it due to many factors ranging from cultural significance to pure curiosity. This observation underscores an important issue in recommender systems that provided ratings have two purposes. On one hand, users want to record their true opinion of a movie. On the other hand, these same ratings are used to generate future recommendations. This setting poses a challenge: how can we accurately predict what users want to watch if we only predict how they’re likely to rate a movie? The discrepancy highlights the need for better evaluation methods that can tell the difference between a user’s interest in watching a movie and their expected enjoyment or rating of it.

4.6.8 Interest vs Popularity

We next investigated the relationship between participants’ interest levels and movie popularity. The total number of ratings in the ML-32M dataset was considered as a measure of popularity. Based on this assumption and the fact that the number of ratings had a very broad range, we first converted the number of ratings to log base 10 values and used it as a measure of popularity. Afterwards, we looked into the overall distribution of popularity as well as the distributions for various interest levels. To get a better understanding of each interest level’s distribution with the overall distribution, we used quantile-quantile (QQ) plots. These plots compare the quantiles of a specific interest category’s log popularity with the quantiles of the overall log popularity. In each QQ plot, the red dashed line represents $y=x$, and it shows identical distributions. Figure 4.9 illustrates the results.

As shown by the results, the least popular movies (those with fewer than approximately 100 ratings) have the highest percentage in the “Not interested” category. As for “Somewhat interested”, the popularity is neutral and shows no significant bias. This can also be seen in its QQ plot which follows the $y=x$ line. Higher interest levels, such as “Extremely interested”, show a decrease in the proportion of low-popularity movies (those with less than 100 ratings). More investigations are needed to establish the root cause of these trends, but we can propose some early hypotheses. One theory is that participants might not be interested in movies with few ratings due to a lack of familiarity. Furthermore, recommendation algorithms might find it hard to suggest less popular movies.

4.6.9 Interest vs Recency of Movies

The other research question with respect to the distribution of variables for each interest level was about the year of recommended movies. Therefore, we looked into this matter by creating a box plot. As shown by figure 4.10, we can conclude that participants were

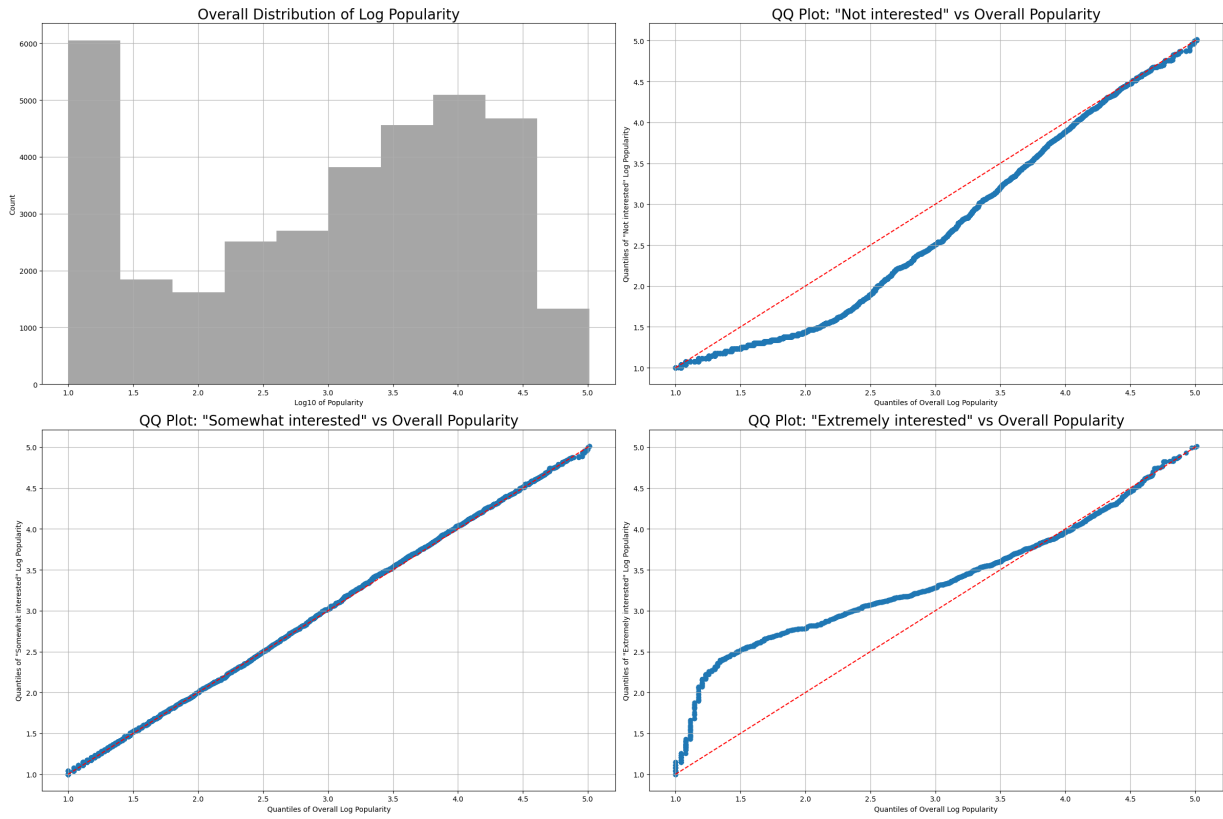


Figure 4.9: Distribution of Interest Levels by Movie Popularity Zone

generally more inclined towards watching older movies, as we can observe that they mostly like the movies from the early 2000s, even though the trend is not very strong and it is rather a general one.

4.6.10 Participant Rate of Judgement

Another important part of our user study was to study the rate at which participants provided their judgments. For this, we first calculated the amount of time spent between viewing a page that included a recommendation and its details and submitting the judgment for that recommendation. Then, these values were averaged for each participant.

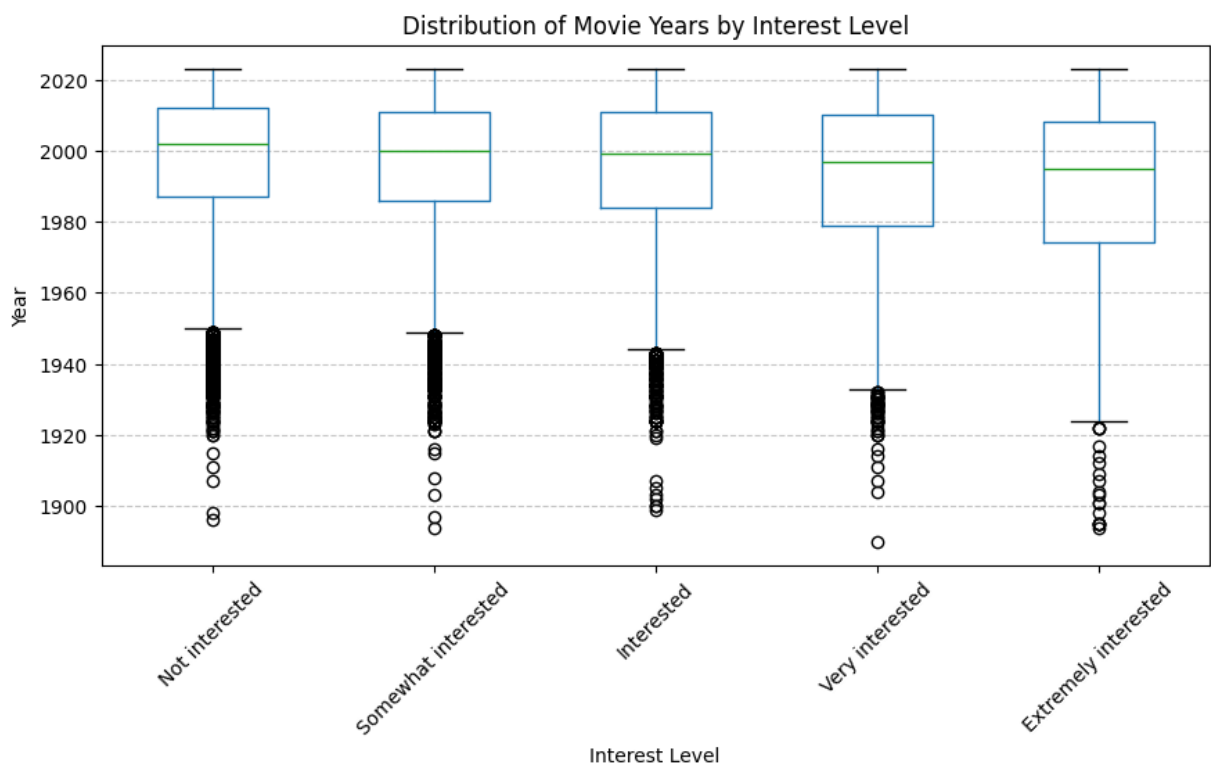


Figure 4.10: Distribution of Movie Years by Interest Level

Our system had a 3 minute timeout, after which participants were automatically logged out if they stayed inactive. Furthermore, participants were instructed that we expected them to judge 2-3 movies per minute. Table 4.23 presents the statistical summary of the average judgement times for all participants.

Table 4.23: Summary Statistics of Average Judgement Time (in seconds)

Metric	Value
Number of Participants	51
Mean Average Time	20.3
Standard Deviation	6.6
Minimum	7.8
25th Percentile	16.3
Median	20.5
75th Percentile	24.8
Maximum	36.0

These findings suggest that our participants generally tended to stick to the expected rate of judgment, although with some variation.

4.6.11 Factors Influencing Interest Level

To conclude our analysis of interest levels, we wanted to assess the overall factors that describe participants' desire to watch movies. Factors included whether or not they have already seen the movie, their level of familiarity, the predicted rating value, and the movie's release year. To answer this question, we utilized a random forest classifier with 100 estimators and an 80/20 train-test split. The random forest classifier, first introduced by Breiman (2001), is an machine learning method that take into account numerous decision

trees to predict the appropriate class for each input data point. For the list of considered features, we chose rating value, release year of the movie, categorical encoding of familiarity level (named *familiarity_int*), boolean value showing whether or not the movie has been seen before by the participant, the rank chosen for the movie during the ranking stage of phase two (used 0 for null values), and whether or not the movie has been used as either phase one or phase two consistency checks (denoted by *p1_check* and *p2_check*).

Figure 4.11 demonstrates the values for feature importance obtained from the trained model. Results show that the predicted rating value and release year are the most crucial factors in determining the interest level. Furthermore, as for the performance of the model, the F1 values for most interest categories were in the range of 0.4 to 0.43. However, for the “Not interested” level, the model had an F1 score of 0.71.

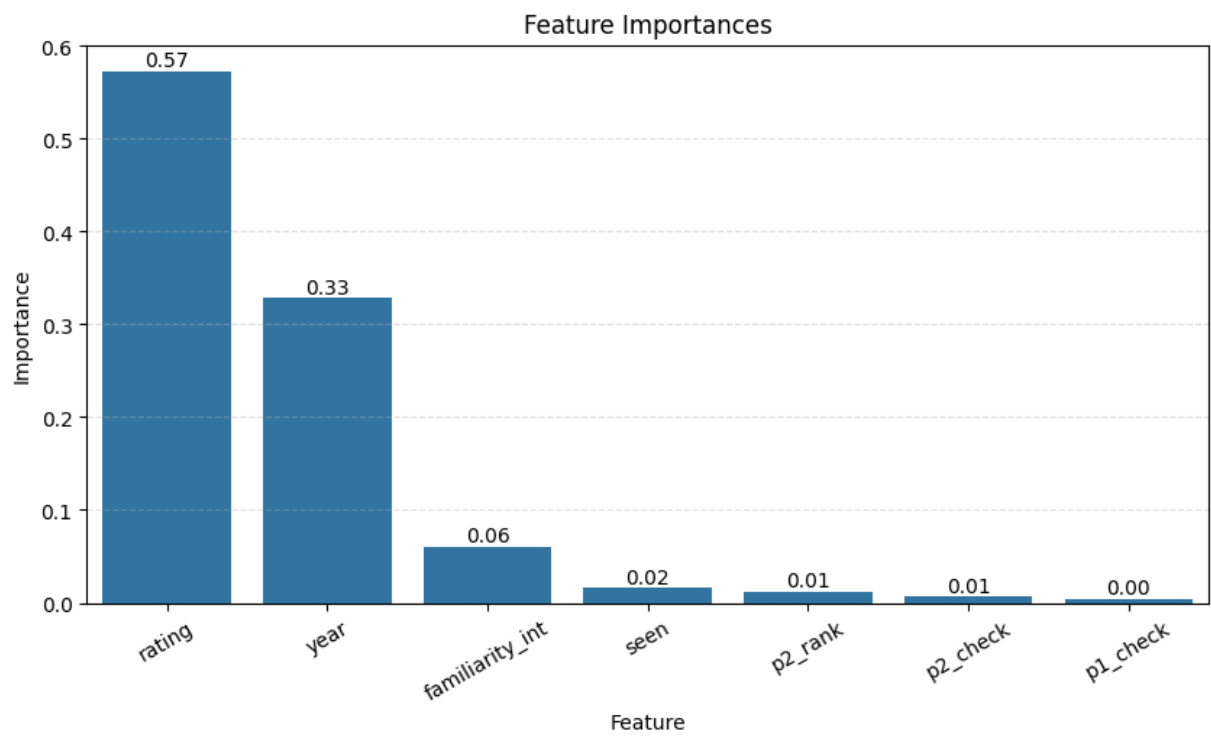


Figure 4.11: Importance of Features in Interest Level Prediction

Chapter 5

Conclusion

In this study, we described the process of creating a novel dataset of movie judgments provided by a subset of MovieLens users. We examined the main differences between recruited participants and average MovieLens users and investigated various factors that influence a person's willingness to watch a movie. For test collection, we gathered judgements from 51 participants on recommendations across three axes: their level of familiarity with the item, their previous or predicted rating based on some provided information, and their interest in watching the movie. Additionally, at the end of each study phase, participants were asked to select their top three most wanted movies, in order of preference, from the titles in which they had the highest level of interest. Our key findings from analysis include:

- Compared to the average MovieLens user, participants had significantly more ratings in their profiles, with the average number of ratings being about seven times higher (1,425 vs. 191).
- Participants showed high test-retest reliability, with a macro-averaged correlation coefficient of 0.81 between their historical ratings and those provided in the study.

- The study revealed a strong correlation between predicted ratings and interest levels for unseen movies, with higher predicted ratings normally connected with greater interest in watching.
- Familiarity with a movie without having seen it increased interest levels. But in general, participants were generally less interested in rewatching movies they had already seen, even if they rated them highly.
- Variables such as predicted rating, movie release year, and popularity were found to influence participants' interest levels in watching a movie.
- Some participants expressed interest in watching movies they expected to rate poorly, mainly for historical, controversial, or cult classic content. This suggests that factors beyond final movie rating can influence viewing decisions.

This thesis has produced a new test collection for recommendation systems evaluation and has provided an analysis of the participants and the collected relevance judgments. Future work will include a comparison of this test collection's evaluation of the various recommendation algorithms compared to traditional recommendation system evaluation methods that make use of a test/train split of user profiles. Additionally, we plan to investigate how effectively each algorithm captures unique user tastes, as observed during the user study. After conducting these analyses, we expect to make the test collection available to other researchers. This will contribute to the broader research community by providing a valuable resource for further studies in recommender systems evaluation and development.

References

- Alam, M. T., S. Ubaid, S. S. Sohail, M. Nadeem, S. Hussain, J. Siddiqui, et al. (2021). Comparative analysis of machine learning based filtering techniques using movielens dataset. *Procedia Computer Science* 194, 210–217.
- Amatriain, X., J. M. Pujol, and N. Oliver (2009). I like it... i like it not: Evaluating user ratings noise in recommender systems. In *International Conference on User Modeling, Adaptation, and Personalization*, pp. 247–258. Springer.
- Beel, J., M. Genzmehr, S. Langer, A. Nürnberger, and B. Gipp (2013). A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In *Proceedings of the international workshop on reproducibility and replication in recommender systems evaluation*, pp. 7–14.
- Beel, J. and S. Langer (2015). A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In *Research and Advanced Technology for Digital Libraries: 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015, Proceedings 19*, pp. 153–168. Springer.
- Bell, R. M., Y. Koren, and C. Volinsky (2010). All together now: A perspective on the netflix prize. *Chance* 23(1), 24–29.

- Bostandjiev, S., J. O'Donovan, and T. Höllerer (2012). Tasteweights: a visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on Recommender systems*, pp. 35–42.
- Bowman, S. R., L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio (2015). Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.
- Breiman, L. (2001). Random forests. *Machine learning* 45, 5–32.
- Burke, R., A. Felfernig, and M. H. Göker (2011). Recommender systems: An overview. *Ai Magazine* 32(3), 13–18.
- Buttcher, S., C. L. Clarke, and G. V. Cormack (2016). *Information retrieval: Implementing and evaluating search engines*. Mit Press.
- Cañamares, R., P. Castells, and A. Moffat (2020). Offline evaluation options for recommender systems. *Information Retrieval Journal* 23(4), 387–410.
- Castells, P. and A. Moffat (2022). Offline recommender system evaluation: Challenges and new directions. *AI magazine* 43(2), 225–238.
- Christakopoulou, K., F. Radlinski, and K. Hofmann (2016). Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 815–824.
- Cleverdon, C. (1967). The cranfield tests on index language devices. In *Aslib proceedings*, Volume 19, pp. 173–194. MCB UP Ltd.
- Dehghani Champiri, Z., A. Asemi, and S. Siti Salwah Binti (2019). Meta-analysis of evaluation methods and metrics used in context-aware scholarly recommender systems. *Knowledge and Information Systems* 61, 1147–1178.

- Dereventsov, A. and A. Bibin (2022). Simulated contextual bandits for personalization tasks from recommendation datasets. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 1–6. IEEE.
- Deshpande, M. and G. Karypis (2004). Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)* 22(1), 143–177.
- Ekstrand, M. D. (2020). Lenskit for python: Next-generation software for recommender systems experiments. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 2999–3006.
- Ekstrand, M. D., J. T. Riedl, J. A. Konstan, et al. (2011). Collaborative filtering recommender systems. *Foundations and Trends® in Human-Computer Interaction* 4(2), 81–173.
- Freyne, J. and S. Berkovsky (2013). Evaluating recommender systems for supportive technologies. In *User Modeling and Adaptation for Daily Routines: Providing Assistance to People with Special Needs*, pp. 195–217. Springer.
- Funk, S. (2006). Netflix update: Try this at home. <https://sifter.org/~simon/journal/20061211.html>. Accessed: 2024-04-21.
- Gondaliya, S. and K. Amin (2021). Multi-batch quasi-newton method with artificial neural network for movie recommendation. *Journal of The Institution of Engineers (India): Series B* 102, 729–742.
- Gunawardana, A., G. Shani, and S. Yogev (2012). Evaluating recommender systems. In *Recommender systems handbook*, pp. 547–601. Springer.
- Han, Z., M. U. Anwaar, S. Arumugaswamy, T. Weber, T. Qiu, H. Shen, Y. Liu, and M. Kleinsteuber (2020). Metapath-and entity-aware graph neural network for recommendation. *CoRR*, abs/2010.11793.

- Harman, D. K. and E. M. Voorhees (2006). Trec: An overview.
- Harper, F. M. and J. A. Konstan (2015). The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5(4), 1–19.
- He, X., L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua (2017). Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pp. 173–182.
- Hidasi, B. and Á. T. Czapp (2023). Widespread flaws in offline evaluation of recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 848–855.
- Hofmann, K., L. Li, F. Radlinski, et al. (2016). Online evaluation for information retrieval. *Foundations and Trends® in Information Retrieval* 10(1), 1–117.
- Hu, Y., Y. Koren, and C. Volinsky (2008). Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*, pp. 263–272. Ieee.
- Jadidinejad, A. H., C. Macdonald, and I. Ounis (2019). Unifying explicit and implicit feedback for rating prediction and ranking recommendation tasks. In *Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval*, pp. 149–156.
- Jannach, D. and C. Bauer (2020). Escaping the mcnamara fallacy: Towards more impactful recommender systems research. *Ai Magazine* 41(4), 79–95.
- Ji, Y., A. Sun, J. Zhang, and C. Li (2022). Do loyal users enjoy better recommendations? understanding recommender accuracy from a time perspective. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 92–97.

- Ji, Y., A. Sun, J. Zhang, and C. Li (2023). A critical study on data leakage in recommender system offline evaluation. *ACM Transactions on Information Systems* 41(3), 1–27.
- Knijnenburg, B. P., N. J. Reijmer, and M. C. Willemsen (2011). Each to his own: how different users call for different interaction methods in recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pp. 141–148.
- Kohavi, R., A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann (2013). Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1168–1176.
- Kohavi, R., R. Longbotham, D. Sommerfield, and R. M. Henne (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery* 18, 140–181.
- Liang, D., R. G. Krishnan, M. D. Hoffman, and T. Jebara (2018). Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*, pp. 689–698.
- Lü, L., M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou (2012). Recommender systems. *Physics reports* 519(1), 1–49.
- Marlin, B., R. S. Zemel, S. Roweis, and M. Slaney (2012). Collaborative filtering and the missing at random assumption. *arXiv preprint arXiv:1206.5267*.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research* 12, 2825–2830.
- Pu, P., L. Chen, and R. Hu (2012). Evaluating recommender systems from the user’s perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction* 22, 317–355.

- Rendle, S., C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme (2012). Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*.
- Rossetti, M., F. Stella, and M. Zanker (2016). Contrasting offline and online results when evaluating recommendation algorithms. In *Proceedings of the 10th ACM conference on recommender systems*, pp. 31–34.
- Roy, A. and S. A. Ludwig (2021). Genre based hybrid filtering for movie recommendation engine. *Journal of Intelligent Information Systems* 56(3), 485–507.
- Said, A., B. Fields, B. J. Jain, and S. Albayrak (2013). User-centric evaluation of a k-furthest neighbor collaborative filtering recommender algorithm. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 1399–1408.
- Sanderson, M. et al. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval* 4(4), 247–375.
- Scholer, F., D. Kelly, and B. Carterette (2016). Information retrieval evaluation using test collections. *Information Retrieval Journal* 19, 225–229.
- Seo, Y.-D., Y.-G. Kim, E. Lee, and H. Kim (2021). Group recommender system based on genre preference focusing on reducing the clustering cost. *Expert Systems with Applications* 183, 115396.
- Shani, G. and A. Gunawardana (2011). Evaluating recommendation systems. *Recommender systems handbook*, 257–297.
- Statistics Canada (2022). 2021 census of population. Census 98-316-X, Statistics Canada, Ottawa, Ontario. Catalogue no. 98-316-X.
- Steck, H. (2019). Embarrassingly shallow autoencoders for sparse data. In *The World Wide Web Conference*, pp. 3251–3257.

- Steck, H., M. Dimakopoulou, N. Riabov, and T. Jebara (2020). Admm slim: Sparse recommendations for many users. In *Proceedings of the 13th international conference on web search and data mining*, pp. 555–563.
- Tasse, G., B. R. Rowe, D. W. Wood, A. N. Link, and D. A. Simoni (2010). Economic impact assessment of nist’s text retrieval conference (trec) program. *National Institute of Standards and Technology, Gaithersburg, Maryland*.
- TensorFlow, D. (2018). Tensorflow. *Site oficial*.
- U.S. Census Bureau (2021). 2020 census of population and housing. Census, U.S. Department of Commerce, Washington, DC. Retrieved [Insert Date].
- Vente, T., M. Ekstrand, and J. Beel (2023). Introducing lenskit-auto, an experimental automated recommender system (autorecsys) toolkit. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 1212–1216.
- Voorhees, E. M. (2001). The philosophy of information retrieval evaluation. In *Workshop of the cross-language evaluation forum for european languages*, pp. 355–370. Springer.
- Voorhees, E. M., D. K. Harman, et al. (2005). *TREC: Experiment and evaluation in information retrieval*, Volume 63. Citeseer.
- Wang, G. and X. Wang (2022). Cold-start sequential recommendation with collaborative boosted meta transitional learning. In *2022 5th International Conference on Data Science and Information Technology (DSIT)*, pp. 1–5. IEEE.
- Wikipedia (2024). Adolf Hitler: The Greatest Story Never Told — Wikipedia, the free encyclopedia. <http://en.wikipedia.org/w/index.php?title=Adolf%20Hitler%3A%20The%20Greatest%20Story%20Never%20Told&oldid=1241331318>. [Online; accessed 28-August-2024].

- Wu, Y., C. DuBois, A. X. Zheng, and M. Ester (2016). Collaborative denoising auto-encoders for top-n recommender systems. In *Proceedings of the ninth ACM international conference on web search and data mining*, pp. 153–162.
- Zangerle, E. and C. Bauer (2022). Evaluating recommender systems: survey and framework. *ACM computing surveys* 55(8), 1–38.
- Zhang, W., Q. Wang, T. Yoshida, and J. Li (2021). Rp-lgmc: Rating prediction based on local and global information with matrix clustering. *Computers & Operations Research* 129, 105228.
- Zhao, W. X., J. Chen, P. Wang, Q. Gu, and J.-R. Wen (2020). Revisiting alternative experimental settings for evaluating top-n item recommendation algorithms. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2329–2332.
- Zhao, W. X., Y. Hou, X. Pan, C. Yang, Z. Zhang, Z. Lin, J. Zhang, S. Bian, J. Tang, W. Sun, et al. (2022). Recbole 2.0: Towards a more up-to-date recommendation library. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 4722–4726.
- Zhao, W. X., Z. Lin, Z. Feng, P. Wang, and J.-R. Wen (2022). A revisiting study of appropriate offline evaluation for top-n recommendation algorithms. *ACM Transactions on Information Systems* 41(2), 1–41.
- Zhou, Y., D. Wilkinson, R. Schreiber, and R. Pan (2008). Large-scale parallel collaborative filtering for the netflix prize. In *Algorithmic Aspects in Information and Management: 4th International Conference, AAIM 2008, Shanghai, China, June 23-25, 2008. Proceedings* 4, pp. 337–348. Springer.

APPENDICES

Appendix A

Web Interface Used for User Study

In this section, we have attached the user study participation guide, which was made accessible to the participants on the web application. Text shown in quotes represent the actual guidelines included in the original document. For instance, below is the introduction:

‘Hello and welcome to the mini tutorial on how to participate in this study. We’ve outlined the steps for each of the two phases with screenshots and brief descriptions. This tutorial follows the same flow you’ll experience during the study. If you encounter any issues with the web app, please contact the student researcher at h2chaman@uwaterloo.ca.’

A.1 Overview and Signup

IRMrec

Evaluation of Recommendation Systems Research Study

Thank you for your interest in our research study. We are a research group at the University of Waterloo led by Prof. Mark Smucker. This study is part of the Master's thesis work of Houmaan Chamani and has been reviewed and received clearance through a University of Waterloo Research Ethics Board.

Goals of the Study:

In order to improve recommendation systems, we need to be able to accurately measure their quality. Our study aims to improve researchers' ability to measure recommendation system quality by collecting from participants, such as you, movie ratings and feedback beyond what you have already provided to Movielens. Our study works by first collecting from you your current Movielens ratings, and then we use your ratings to generate recommendations from a large number of different recommender algorithms. We then have you rate and give feedback on these recommendations so that we can measure the quality of these and future recommendation algorithms. The data we collect will form a long term resource available to researchers to advance the study of recommendation systems.

Summary of Participation Requirements and Remuneration:

- To participate in the study, you must be 18 or older and a resident of the United States or Canada.
- The study has two phases. The first phase will take you about 2 hours of time. The second phase is optional and may take up to an additional 5 hours of time.
- To thank you for participating in the first phase, we will give you an Amazon e-gift card of \$40 CAD or \$30 US depending on whether you are a resident of the US or Canada.
- If you participate in the second phase of the study, we will give you an Amazon e-gift card worth \$15 US or \$20 CAD per hour.

Expression of Interest:

If you are interested in participating in our study, please submit your information below, and we'll be in touch shortly with further details. Submission of your information below does not oblige you to participate nor does it give your consent, i.e. you are merely expressing your interest in the study and requesting further information.

Full Name:

Email Address:

Figure A.1: First part of the overview page

Summary of Participation Requirements and Remuneration:

- To participate in the study, you must be 18 or older and a resident of the United States or Canada.
- The study has two phases. The first phase will take you about 2 hours of time. The second phase is optional and may take up to an additional 5 hours of time.
- To thank you for participating in the first phase, we will give you an Amazon e-gift card of \$40 CAD or \$30 US depending on whether you are a resident of the US or Canada.
- If you participate in the second phase of the study, we will give you an Amazon e-gift card worth \$15 US or \$20 CAD per hour.

Expression of Interest:
If you are interested in participating in our study, please submit your information below, and we'll be in touch shortly with further details. Submission of your information below does not oblige you to participate nor does it give your consent, i.e. you are merely expressing your interest in the study and requesting further information.

Full Name:

Email Address:

Age:

Country of Residence:

Contact Information:

Student Investigator: [Houmaan Chamani](#)

Faculty Supervisor: [Prof. Mark D. Smucker](#)

Figure A.2: Second part of the overview page

A.2 Login

Login to IRMrec

Username:

Password:

Figure A.3: Login page

A.3 Homepage

‘After logging in, you’ll land on the homepage. Throughout the study, you’ll find a drop-down menu on the top right side of the page. The homepage provides a summary of the study steps. To start the main task, which involves judging recommended movies, click on the blue button or select the second item from the drop-down menu.’

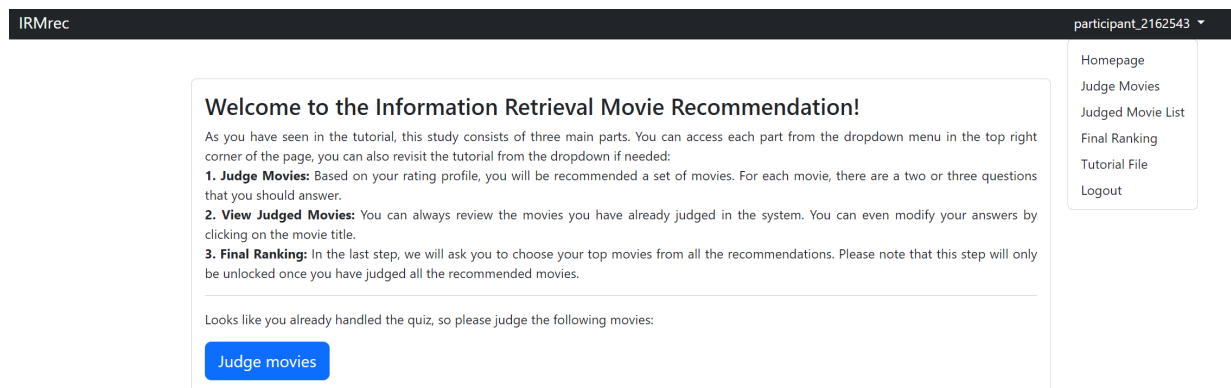


Figure A.4: Home Page

A.4 Movie Judging

‘To judge a movie, click on the drop-down boxes to make your choices and hit the blue submit button when you’re finished. The questions are as follow:

A - How familiar are you with this movie? There are 13 options to choose from for the answer, divided into two groups.

1. Have not seen the movie

- 1.1. **Never heard of it:** You have not seen the movie and have never heard about it.

1.2. **Familiar with the movie:** Still unseen, but you have heard about it from somewhere.

1.3. **Very familiar** (read reviews, seen trailers, etc.): Still unseen, but you are very familiar with the movie, for maybe you’ve watched some trailers or read reviews of the movie or have even done a search about the plot, etc.

2. **Have seen the movie - 0.5 to 5 stars**

2.1. Rate the movie based on your experience, using the same rating scale as MovieLens (0.5 to 5 stars). In some cases, you may be given a movie to rate that you’ve already rated on MovieLens; please rate them again for this research study. These ratings are for the research study only and will not be associated with MovieLens.

B - How interested are you in watching this movie via a streaming service?

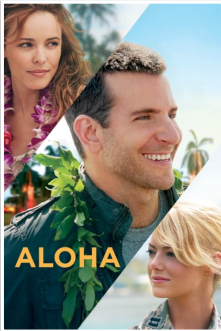
Imagine you are given a subscription for a streaming service for free. In that case, what would be your level of interest in watching this movie on that platform, assuming the movie was available to watch? Please note that this holds for both seen and unseen movies. For example, for some seen movies, you may be very interested in watching them again, while other seen movies, even ones you love, you might not care to see again. There are five answers to choose from for this: Not interested / Somewhat interested / Interested / Very interested / Extremely interested Your interest level should correspond to the likely order that you’d watch the movies. For example, you’d watch the “extremely interested” movies before you’d watch the “very interested” movies and so forth. You would never watch the “not interested” movies.

C - For unseen movies only: if you were to watch this movie, what would you predict your rating of it to be? For this question, you are to predict what your MovieLens rating would be for the movie after you watch it. You will only answer this

question for unseen movies. This question is disabled for any movies you’ve already seen as per the first question.’

IRMrec

participant_2162543



Aloha

A celebrated military contractor returns to the site of his greatest career triumphs and re-connects with a long-ago love while unexpectedly falling for the hard-charging Air Force watchdog assigned to him.

Release Year: 2015 **Runtime(min):** 105

Cast: Bradley Cooper, Emma Stone, Rachel McAdams, Bill Murray, John Krasinski

Director(s): Cameron Crowe

Genres: Drama, Comedy, Romance

Languages: English, Русский

How familiar are you with this movie?

Choose an option

How interested are you in watching this movie via a streaming service?

Choose an option

For unseen movies only: If you were to watch this movie, what would you predict your rating of it to be?

Choose an option

Submit

Figure A.5: Default view of the movie judgement page

A.4.1 Judging Speed and Use of other Websites

‘It is important to work efficiently when judging movies. While we want you to judge movies carefully, you should not dwell for a long time on your decisions. On average, we expect you to judge movies at a pace of approximately 2 to 3 movies per minute. This is an average pace. For some movies you’ll have quick decisions, and for others you’ll need to contemplate the plot summary, actors, director, etc.

You should not make use of other websites for additional information about the movie. Please use the information on the page to guide your decision. There is no time budgeted to research a movie beyond the information that we provide.’

A.5 List of Judged Movies

‘At any point during the study, you can view the list of movies you’ve already judged by selecting the corresponding option from the top right dropdown menu. On this page, you’ll see the number of remaining movies to judge. By clicking on the blue text in the middle of the page, you can resume the judgment task. If you wish to edit your answers for previously judged movies, click on the movie’s title. You can also sort the list based on your rating or willingness to watch using the sorting option in the top right section.’

IRMrec

participant_2162543 ▾

Judged List		3 Movies Left to Judge, Click to Resume		Sort Movies By: ▾	
	Title (click to edit movie rating)	Year	Familiarity	Your Rating	Willingness to Watch
	Aloha	2015	Seen	3.5	Very interested

Figure A.6: List of judged movies

A.6 Judgement Completion

‘Once you’ve judged all the recommended movies, you’ll reach this page. Here, you can choose to review your choices or make changes to any of them. Once you are satisfied with your judgments, proceed to the final step, which involves ranking the movies you are most interested in watching.’

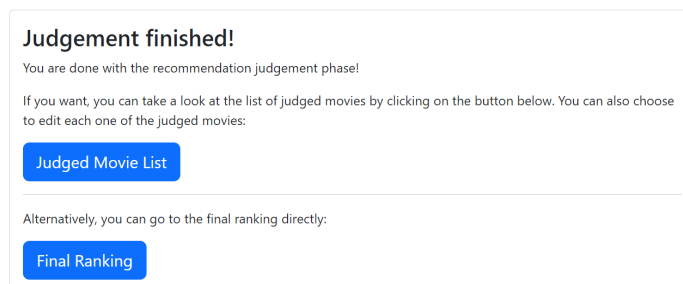


Figure A.7: Judge completion page

A.7 Ranking

‘In this page, you can see the list of all the top movies based on your previous answers for the judgment task. In this part, we ask that you choose your top three items out of this list and rank them. You can imagine the question in the following way: **if you were to pick three movies out of this list, which three would you pick and in what order would you watch them?**. We call this the watch queue. As you can see in the picture above, you can choose to include each item in your watch queue. Please note that the queue is updated as you click on new items. For instance, if you have chosen three movies such as M1, M2, and M3 and ranked them, should you choose to place another movie (like M4) in the first rank, the previous ranking would change and your new watch queue would be M4, M1, M2. Also, you can choose to empty a previously filled slot in the watch queue using the trash can icon next to each slot. Please note that a single movie can not be in the watch queue twice, and the three items should be **unique, and all three slots should be filled with a movie**. Once you have your queue built to your satisfaction, please submit your watch queue using the submit button.’

IRMrecparticipant_2162543

Choose Your Top Picks

	Title	Year	Place in Top Three
	Jumanji	1995	1 2 3
	Heat	1995	1 2 3
	Casino	1995	1 2 3
	Get Shorty	1995	1 2 3
	Leaving Las Vegas	1995	1 2 3

Watch Queue

Rank 1
- 

Rank 2
- 

Rank 3
- 

Submit Queue

Figure A.8: Movie ranking page

A.8 Ranking Unavailable

‘Please note that ranking and creating your watch queue is the final step. Ensure that you’ve judged all recommended movies before attempting to rank them. If you click on the ranking option in the top right dropdown menu before completing the judgments, you’ll see this page. You can resume judging movies by clicking the first button on the left (Judge movies) or view your previously judged movies using the second button on the right (View judged movies).’

Final Ranking Unavailable!
 You still have remaining movies to judge before proceeding to final ranking:

[Judge movies](#)
[View judged movies](#)

Figure A.9: Ranking page unavailable

A.9 Post study questionnaire

‘When you submit your watch queue, you will be redirected to this page. This is the last step of each phase. We’d like to know your opinion on any part of the study including the user experience, recommended movies, goal, basically anything that comes to your mind. There is a text box to type in your feedback.’

IRMrec

participant_1162542 ▾

Almost there!
 The only remaining part is the post-study feedback for phase one!
 Please give us any sort of feedback you can think of regarding the system, user experience, study design, or any other aspect! We will use your valuable response to improve future studies.

Characters left: 300

Submit

Figure A.10: Post study questionnaire

A.10 Final page

‘This is the last page you will encounter at the end of phase one. As indicated, to receive reimbursement for your time commitment, please send an email to the student investigator using the same email you signed up with and also make sure to include your unique participant ID in the email. Lastly, if you’d like to also participate in the second phase, which involves judging more recommendations, please also indicate this in your email. We will gather all responses from the first phase and contact selected individuals to take part in the second phase.’

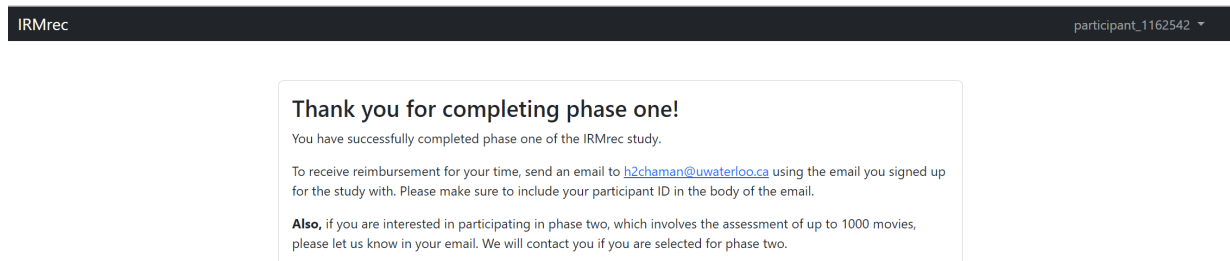


Figure A.11: Final web application page

Appendix B

Web Interface Quiz

The following quiz was presented to participants before they could proceed with the study. Participants were required to answer all questions correctly to continue. Correct answers are marked with an asterisk (*).

1. When judging movies and specifying your interest in watching the movie via a streaming service (Netflix, etc.), what should your interest level correspond to?
 - The degree to which you like the movie or anticipate liking the movie, i.e. your rating.
 - *The likely order that you'd watch the movies, i.e. you want to watch the movies you are more interested in before the movies you are less interested in.
2. When judging movies and specifying your interest in watching movies via a streaming service (Netflix, etc.), can you be interested in watching movies that you've already seen?
 - *Yes. We all know the joy of finding a favorite movie available on a streaming service and wanting to watch it again.

- No. If you've already seen a movie, you should not be interested in watching it again.
3. When judging a movie and specifying your interest in watching a movie via a streaming service (Netflix, etc.), you have seen the movie previously and rated it 5 stars, the highest rating possible, can you be uninterested in watching it again?
- *Yes. It is possible to love a movie but not be interested in watching it again.
 - No. If you love a movie, you must be interested in watching it again very soon.
4. When judging movies, for the movies you've already seen, what is the rating scale that you will use?
- *The Movielens rating scale of 0.5 to 5 stars, and I should rate the movies the same as if I was rating them on Movielens.
 - On a scale of 0 to 100 where scores over 50 are for movies I like and scores less than 50 are for movies that I dislike.
 - There is not a rating scale. I am supposed to describe in words my opinion of the movie.
5. For a movie you have not yet seen (an unseen movie), what are the possible judgments?
- *1. Never heard of it. 2. Familiar with it. 3. Very familiar with it.
 - 1. Inappropriate or bad recommendation. 2. Okay recommendation. 3. Good recommendation.
 - None of the above.

6. For an unseen movie, we will ask you to predict what your Movielens rating would be for the movie after you watch it. Is this the same as your interest in watching the movie?
- Yes. The higher your predicted rating, the more interested you are in watching the movie.
 - *No. Your predicted rating is based on how much you think you'll like the movie, but your "interest in watching" level is based on your true desire of what order you want to watch movies in.
7. How can you view the list of movies you've already judged during the judgment phase?
- It is not possible
 - Refresh the homepage
 - *Click on the corresponding option in the top-right dropdown menu
 - You will automatically be redirected to that page after judging each movie
8. How can you make edits to your answers for a movie you have already judged?
- A page will pop up right after submitting the first set of answers for that specific movie
 - *By clicking on the movie title available on the judged movie list page
 - It is not possible
9. Where can you find the number of remaining movies to be judged?
- Go to the Homepage
 - You will see it after each movie you judge on the "Judge Movies" page.

- *If you go to the "Judged Movie List" to review your judgments, you will see the number of remaining movies to be judged, too.
10. How can you access the ranking page before finishing all judgment tasks?
- *You can't as it will become available once all movies have been judged
 - By Clicking on the corresponding option in the top-right dropdown menu
 - You can access it via a link on the Judged Movie List page
11. What is the most accurate description of the ranking phase?
- Randomly fill out the available spots from the pool of shown movies on the page
 - *From all the shown movies, pick the three you are most interested in watching and rank them in order based on your interest to watch them, i.e. which movie you want to watch first, second, and third.
 - Choose any three movies and rank them.
12. When judging movies, how fast should you work?
- I should be very deliberative and take as much time as needed to produce the most accurate judgments.
 - *I should judge at a pace of about 2-3 movies per minute. As such, I can only use the information provided to make a decision. I need to balance accuracy with speed.
 - The faster I work the better regardless of my accuracy.

Appendix C

Participant Feedback

In this section, we have included all the feedback received from our selected participants (SP) during the user study. Each row of the table is a participant.

First Phase Feedback	Second Phase Feedback
Unclear if initial quiz was referring to Movie-lens interface	-
-	Time to login again should be extend up from 3 minutes. Love to see the rating of the movie. If I felt a movie skewed much younger I would be less likely to watch it.
I found it all very easy to use!	Nothing I can think of. It all worked well and intuitively!
Maybe a wider range of choices for the ranking stage.	Final Ranking could be expanded upon in future studies perhaps

-	Actors and directors are a huge factor in swaying peoples' perceptions of movies. Maybe it would provide some more insight if you included a "Never heard of it - familiar with cast/director" option, or something to that effect.
None, the system was great, I had no issues with it.	Experience and design was great! No complaints or feedback.
The instructions should be put on a page before the questions about the instructions.	No new feedback to give.
It has been fine so far, gets a little repetitive but there's nothing you can do about that really	Same as last time, got repetitive but there's nothing you can do about it
I didn't quite know what to do what multiple films from series I had not seen. I am less interested in watching all subsequent films than the first, but I'm not always sure which one is the first.	No additional feedback from last time!
When first logging in to review the updated consent forms and answer the quiz, I did get timed a couple times. Interface for rating movies was straight forward and east to use.	After completing phase two, I chose to edit a few selections from my judged movies list. I think it would be helpful if I was brought back to the screen with all my judged movies, instead of having to click another button to see my full list.

I enjoyed the survey, there were several movies I never heard of that I'm now interested to see	Maybe you should be able to opt out of certain movie genres. For example I have no interest in most animated movies so I will always rate them lower.
I find previews the most useful predictor of whether I will like a movie. When I based movies on reputation and descriptions, I liked about 10% of the movies I chose to watch. When I use previews to make my selections, I like about 60% of the movies I chose to watch. There's also a lot to be said for "made by the same writer (or director)".	Have to keep in mind, even though I agree 2.5 stars is not very good, it's an expected average based on my existing data. I pessimistically don't expect many to be better than that. Sturgeon's revelation. 2.5 as an expected average after I have already been very selective. Judging these synopses, how being formulaic is safe, but hard to be great. I had very little information to make these interesting to me. Previews are much more informative, but time consuming. Inflating based on past experiences, deflating based on genres I am not interested in: Horror, Stand Up.

This was really enjoyable. There are a couple movies that I saw so long ago, I'm slightly not sure I have seen them, but that was only a very few. Product of having seeing a lot of films and being over 50, I guess! The instructions were more confusing than the actual ranking but that's how things often go. The quiz is good idea. Hope to be part of the next phase!	This was a really easy system to use and exposed me to a lot of movies I'd like to add to my list. I appreciated being able to change in the moment before I submitted my rankings and also, though I didn't use it, knowing I could edit if I wanted to. I also liked being able to work at my own pace and time as it was very easy to log in and out. Happy to be a survey subject for you anytime!
I found the descriptions really helpful in making decisions. I really like the interpretation of ranking (like on movielense). I wish you displayed all the options with instead of a drop down. It would have reduced the number of clicks in half! I really struggled with rating the unseen movies that I don't intend to watch. The most important determinant of rating for me is quality, and I can't tell that without more info (and potentially watching it myself). This was particularly relevant for the Holocaust movies because I have Opinions. I'm surprised you didn't ask what languages I speak.	-

There was one movie that had no description. I was still able to make a judgement because I knew what Joseph Campbell's hero's journey was, otherwise I would have been completely uninterested.	After finishing judging movies in phase one of the study, all of the movies were fresh in my mind when doing my final ranking. When doing the final ranking for phase 2, since I did it over multiple days, I found myself struggling to recall some of the movies I was completely unfamiliar with but showed interest in seeing. Putting some or all of the information used to make my judgements would have helped me with recalling what each movie was about.
1x-2x I was asked to judge a movie and then later the original movie that inspired the remake. Because the movies have the same plot, my interest level & familiarity were practically identical (apart from not knowing it the original or remake). This may be the case for most people, and avoiding duplicates may help improve the data &/ recommendations. Several movies were simply redone versions of other movies or TV series. In this case I had seen all of the original series and loved them, and while I would watch them, I would infinitely prefer to find a new/original movie.	Separate final queues for watched & unwatched may provide more info; I can easily rank limited faves &/ unknowns, but putting them in the same queue is like apples & oranges. Sometimes I was frustrated to see just a few poor movies from a particular series or director: I wish I could have left a note of the good ones (e.g. Phantasm, Mike Flanagan, &c). Twice I thought I knew why the algorithm picked a particular suggestion (e.g. The Foundling (1939) because of Paper Moon (1973)), which made the next few q's more about resisting gaming the system. Unsure if it matters or how to prevent.

I had no problems with any phase of the study. It was enjoyable and easy to understand.	A progress bar might be helpful to let people know how far along they are in the study / survey.
It was pretty easy & simple to get through. I was a little surprised with how many of the recommendations I had never even heard of.	Wow, that was a lot of movies to sift through haha. Once again, very surprised at how many I totally hadn't heard of, whether because they are new, very old, or foreign films.
quite a smooth experience, had no problems at all, got a lot of movies I was interested in seeing.	alot of movies i've never heard of, I still enjoyed doing the study :)
I had multiple short features and comedy specials amongs the movies. I think it might bias the study since they aren't movies per se.	Maybe a "previous" button could be beneficial when rating, in case you clicked submit too fast. Since we are not able to access the movies in the judged list by most recently judged, I can become tedious to find the last rated movie.
I really don't have much to say. the process was pretty straightforward	The only issue I faced here was putting together my final to-watch queue. I had to minimize the zoom on my page to find the submit button. I tried a few times and couldn't figure out how to submit my choices. For reference I'm using firefox.
Very straight forward study and design. Rating movies you haven't seen nor heard of is very difficult so possibly remove that option as it could skew the results.	Same as phase one. Very easy to use the interface and overall value seems in the right place. Still not sure the value of trying to guess my rating on unknown movies.

You should be able to pause to go to the bathroom or get a drink. Also, I hate animated films, documentaries, and most foreign films but that is mostly what you gave me to rank.	-
That was fun. I was surprised how many short films were shown. I find it curious you ask me to predict what I'd rate movies I've never even heard of.	Got tired of seeing "movies" for stand-up comedian performances. Not interested in those at all. I'm looking forward to watching some of the recommendations. If nothing else, this was a good way to be exposed to lots of movies I've never heard of.
I thought everything was very smooth! The only criticism, some of the titles I was being asked to judge weren't really movies. Many were short films that run shorter than 10 mins, I'm unsure of whether those titles necessarily belonged in this sort of study. Thanks!	-
Everything was good. Easy to understand. The begin questionnaire was a little confusing. The questions themselves were difficult to understand what you were looking for on some of them. Besides that it was very easy to understand.	Everything was good. Easy to understand. Being able to start and stop was really good. No complaints.
System was good.	-

Overall, the system seemed well designed. A back button might be nice, since a couple of times while judging I had to go back to double check a response I'd given.	Some sort of indicator of the remaining number of movies while on the rating screen might have been useful so I wouldn't have to go to the Judged Movies List to check. (I was trying to do a certain number each day to break it up.)
Overall, this was an enjoyable study to take part in. The tutorial was very helpful in conceptualizing the difference between interest and rating. I did run into a UI bug in the final ranking. When using Chrome browser, the 'submit' button is not visible, even when scrolled all the way to the bottom of the page. In Edge, it is visible, but it is pushed down almost out of field of view if your selected movies have longer titles. I was barely able to access the button to submit, even when scrolled all the way to the bottom of the page. Other than that, though, the site was very easy to navigate.	Nothing new comes to mind beyond what I mentioned at the end of phase 1. I was glad to be a part of this study!
2-3 minutes might be to long. If you are familiar with the movie, you can answer in only a few seconds. If you are unfamiliar with the movie, reading the description takes less than a minute.	I am a frequent user of Movielens, and the recommendations from it have highly effected my responses. I do not know if this is an issue for your study, but I thought it could be worth noting.

1. A handful of the movies presented to judge were ones I knew I'd seen and already rated on ML - and I wondered if that was by design. 2. A handful of "" "" were ones I'd seen but not yet rated on ML. Now I don't remember which ones they were and "Judge Movies List" is no longer available! Can the list be made available to those finished Phase1, or emailed? 3. A handful of "" "" were either tv series, tv miniseries, movie shorts or cartoon shorts - ie: not movies, an ongoing problem/issue with the ML database. These are often tagged as such by users. Is there a way to filter them out by tag?	I found my participation worthwhile: as a movie nerd, I try to keep track of watched films & ratings but sometimes, particularly if I want to think about the film a bit before rating, I forget to add them to ML or IMDb — and a few movies presented for me to judge were ones that had slipped through, so it was good to be reminded of them. Same for films I've seen but wouldn't mind seeing again. I added my final JML to a spreadsheet so I can filter the ones I want to add to my watchlist. Now, if only there was a way to export my JML 'Seen' scores from this study directly into ML...
The only thing I can think of that would have made the experience more pleasant would be: 1. Splitting the instructions and quiz questions up, so you read a section of instructions, then answer a few questions, and so on. 2. Some kind of break from rating movies. So after every 20 movies, getting a little animation or words of encouragement or something.	Thank you for the opportunity to participate!
I thought the UI was easy to use and I found some very interesting recommendations during the study that I will definitely be adding to my watch list.	My only suggestion is that it would be nice to be able to see some indication of how much progress you have made on the ratings (like a progress bar or percentage on the screen).

The submit button in Ranking - Top Picks was not visible until I made the window fullscreen (F11) and then it appeared. No other comments at this time.	-
Being able to exclude certain categories of movies from the recommendation system would be ideal. For example, I am generally not a fan of and not interested in documentaries of any kind, or similarly I am not interested in "behind the scenes" types of films.	Giving the synopsis of the movies under the Ranking List would help determine which order I'd like to watch movies in. I looked through so many over so many days that I cannot remember the synopsis of some of the movies I was extremely interested in watching, and some definitely have a higher priority than others.
Some of the language selections on recent Hollywood films seem to be incorrect. I also feel like the definition of 'movie' was stretched to include several short films, TV shows, and standup comedies, which don't seem directly comparable to feature length films - short films in particular are such a different viewing experience.	Still have issues with the comparability of TV movies/TV shows/shorts, and some of the foreign language films (particularly the Russian ones) had no useful information in the descriptions.

Adding a section for the judge to explain how they capture, curate and consume media. I think this drives my ranking of movies to watch next. Also, the sources of visibility on algo ranking would be helpful directing me on which movie ranking I should choose. I love your service by the way. I love that you allow for a data dump on movies watched. Very nice job you guys!	I'd be curious on what insights you gain from our reviews and ranking. Any summary would from your research would be greatly appreciated to provide context on this study. Thank you again.
Liked it! Would like to have been able to fill out a watch queue more than once, with additional movies thrown in. I would also have liked to rank all of them on a "watch queue". More generally on the movie judging section, I would have liked to see a skip button or something that says I don't normally watch these types of movies, making them fairly hard to rank. And I assume this is outside the scope, but a quick trailer really would have reinforced my rating on some of these movies, trailers would have made it easier. Thanks for inviting me!	I would like to rate more things for the final watch queue or a larger amount of queues to create?

I always find it difficult to assess what I would think of a movie I haven't seen. For movies I've already rated on Movie Lens, I put in the same rating. How long ago I saw a movie is a big factor in how interested I am in seeing it again, as is the rating and the type of movie. Also, a significant portion of the data seems to include music/concert videos, animated shorts and other short films, and most of all comedy specials. I don't think any of that belongs in this data, personally. I hid anything like that if it floats to the top of my recommendations in movielens.	I think my feedback is similar to that of phase 1. A lot of short films, documentaries, comedy specials, etc, that don't quite fit with standard films in my opinion.
Everything worked smoothly. I think that I've got a feel for what you've presented and the intention, but there's nothing to stop me from rating films - seen or unseen - which I do freely and with more than a little gusto.	The interface was easy to use. The only slightly irritating part - also a problem with Movielens - is that a large number of the "films" were episodes of television series, 5 or 6 minute Bugs Bunny cartoons, and a host of others that are certainly not feature films or "movies" at all. The last time I looked the TV series Firefly was still listed as a film.

I have found it hard to predict scores for movies that I don't know or am not interested in watching. Especially, movies that are not really movies, like comedy shows, TV episodes, shorts. Normally, I would just leave them out of my rankings, unless - they were really great. One things to improve: short auto logout time. I got logged out when I tried to write and send this note. Good that I made a copy before sending "Submit", otherwise it would be lost.	-
You might save time by asking first if there is any type of movie that the person is going to automatically rank "Not interested" every time. Like for example, I'm going to put "not interested" on any animated movie or any movie that I think is going to "preach" to me.	Let the person rule out a whole subset of movies that they know they will rank as Awful. Like for example, for me that would be Asian movies or animated movies or movies that are too "preachy".
Having a quiz was surprising. Usually I figure things out as I go along rather than memorize instructions. The quiz made me pay more attention. The set-up was clear.	Maybe because I've already rated a lot of movies, I ended up having to judge cartoons and short films. I don't think they are on a par with full-length movies so I would omit these from the study if you can.

This was interesting to do. Felt a little unfair but a lot of movies were basically disqualified because I usually don't/won't watch foreign films– not strictly, but it's rare that I will watch a foreign film. Similarly I'm not as interested in films more than 40 or 50 years old. I feel guilty judging films based on that criteria but it's honest.	Thank you! Enjoyed reading about the movies, and have a useful resource for when I need something to watch!
Perhaps have an option for movies you've seen but don't really remember - hard to rank? Or perhaps instructions that if you've seen it but don't remember it enough to give it a ranking, then designate as "unseen".	It would be interesting to get a file of the moves we rated - a lot of foreign films I've never heard of! I mentioned in the first part, but having an option or instructions on what to do if you have seen one but don't remember enough to rate it. Thank you!

I was confused for a bit answering the questions on the instructions because I skipped past the first introductory line without reading it. I eventually scrolled back up to the top, but maybe the first question should start with "Based on your reading of this study's instructions (posted above),..." I did figure this out in the end though. Everything else was clear to me. I did find there were a lot of documentaries on music, which I have no interest in watching.	Movies I ranked near the start of the study and ones I ranked near the end seemed to have some duplicates. I tried not to let it affect me, but the mood I was in probably affected how receptive and adventurous I was in my movie tastes. A big factor for my likelihood to watch a movie sooner is whether or not it is in English. I watch movies with subtitles, but since English language movies are easier to watch, because you can glance away, I am always more likely to watch those. Also, while this was a good system, there is no way to find the most recent movie I ranked to edit it.
--	---