User Perception of Different Recommender Algorithms in the Book Domain

Bachelor's thesis

Submitted in Partial Fulfillment of the
Requirements for the Degree of
Bachelor of Science (B.Sc.)
in the field of computer science

by

Clemens Steinmann

First reviewer: Prof. Dr. Frank Hopfgartner
Second reviewer: Dr. Ing. Stefania Zourlidou

Koblenz, December 2023

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

	Ja	Nein
Mit der Einstellung dieser Arbeit in die Bibliothek bin ich einverstanden.	☑	☐

Hachenburg, 08.12.2023 Csteinmann

(Ort, Datum) (Unterschrift)

Abstract

Developments in recommendation system research suggest that objective metrics may not be the best way to measure user perception of recommender systems. We applied methods, proposed in a study by Ekstrand et al., to the book domain to investigate how users perceive different recommendation lists generated by three widely used recommendation algorithms. The factors under consideration are accuracy, diversity, degree of personalisation, satisfaction, and novelty. Our results largely align with those of previous studies, with novelty and satisfaction influencing the initial impression of recommendation lists negatively and positively, respectively. Furthermore, a positive influence of diversity on user satisfaction was observed. We expanded the investigation of user perception to include the incorporation of demographic data. We determined an expert score representing the user's domain expertise, which had a significantly positive impact on user satisfaction. Additionally, we compared the subjectively perceived characteristics of the algorithms by users with objective metrics measuring the same properties.

Zusammenfassung

Entwicklungen in der Forschung zu Empfehlungssystemen deuten darauf hin, dass objektive Metriken möglicherweise nicht der beste Weg sind, um die Nutzerwahrnehmung von Empfehlungssystemen zu messen. Wir haben Methoden angewendet, die in einer Studie von Ekstrand et al. vorgeschlagen wurden, um im Buchbereich zu untersuchen, wie Nutzer unterschiedliche Empfehlungslisten wahrnehmen, die von drei weit verbreiteten Empfehlungsalgorithmen generiert wurden. Die berücksichtigten Faktoren sind Genauigkeit, Vielfalt, Grad der Personalisierung, Zufriedenheit und Neuheit. Unsere Ergebnisse stimmen weitgehend mit denen früherer Studien überein, wobei Neuheit und Zufriedenheit den ersten Eindruck von Empfehlungslisten negativ bzw. positiv beeinflussen. Darüber hinaus wurde eine positive Wirkung von Vielfalt auf die Nutzerzufriedenheit beobachtet. Wir haben die Untersuchung der Nutzerwahrnehmung erweitert, um die Einbeziehung von demografischen Daten einzuschließen. Wir haben einen Expertenwert ermittelt, der die Domänenkenntnisse des Nutzers repräsentiert und einen signifikant positiven Einfluss auf die Nutzerzufriedenheit hatte. Zusätzlich haben wir die von den Nutzern subjektiv wahrgenommenen Eigenschaften der Algorithmen mit objektiven Metriken verglichen, die dieselben Eigenschaften messen.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In today's digital age, recommender systems play an important role in assisting users with finding content, products, and services that match their personal preferences. These systems are not limited to a single domain but are used in various areas, from online shopping [79] to entertainment [81, 30], and news [1, 10]. As digital content continues to grow rapidly, users are increasingly relying on recommender systems to help navigate through the overwhelming amount of information and receive personalised recommendations [8, 67]. As the usage of recommender systems grows, the more likely it is to experience the unwanted effects of a filter bubble [64]. Since there have been many studies investigating the technical aspects of recommender algorithms, this study takes a user-centric approach, aiming to understand how individuals perceive and interact with these recommendation systems, with a particular focus on book recommendations.

The way people discover and enjoy books has significantly changed with the arrival of digital technology [97]. Readers now turn to online platforms and recommender systems to discover their next favourite book. As Lee Shippey said: 'The right book at the right time, may mean more in a person's life than anything else', books hold a unique place as a source of knowledge and entertainment [83], making book recommenders a crucial subdomain of recommender systems. Customised book recommendations can greatly enhance the reading experience, helping users discover new books and experience more diverse and fulfilling reading journeys [54]. This study has the potential to drive innovation in the design and implementation of these systems, benefiting not only academia by adding knowledge in the field of recommendation systems and user perception, but also practical applications by contributing to the improvement of the design and optimisation of book recommendation algorithms, thereby benefiting platforms, publishers, and readers.

However, book recommendations pose unique challenges. Literature is inherently complex, and readers' preferences are multifaceted and nuanced, making it difficult to capture the essence of a book [92]. What resonates with one reader may not resonate with another. Each book is a unique combination of themes, genres, writing styles, and cultural context. Furthermore, readers are often not stuck to a single favourite genre and have rather many genres they enjoy reading. The ever-expanding array of books across various genres increases these challenges.

Current book recommendation approaches like those employed by Amazon [4] and Goodreads [31], provide insight into how recommendations can be presented but also that they more often lack personalisation. Amazon often recommends books based on purchases by similar customers or trivial connections, such as belonging to the same book series, but these recommendations can be commercially driven since vendors can sponsor their articles to be shown more often [32, 53]. Goodreads, on the other hand, relies on genre-based recommendations, suggesting books enjoyed by other readers or matching recommendations with the user's preferred genres [11, 19].

In this context, user perception of recommender algorithms is shaped by factors such as recommendation accuracy, diversity, trustworthiness, novelty, and satisfaction [26]. This research aims to investigate these aspects and their impact on the effectiveness and the users' perception of book recommender algorithms. Additionally, the presence of bias in recommendations can affect fairness, diversity, and accuracy, which, in turn, influences how users perceive these systems [27, 24]. This research aims to investigate user perception as well as consider the potential biases in user groups and their influences on the perception of recommender algorithms.

Despite the progress in understanding the technical aspects of recommender systems, there remains a research gap regarding user perception [69]. Less attention has been given to understanding how users actually feel and respond to the recommendations they receive [39]. Therefore, this study aims to fill this gap by examining whether and to what extent the results of previous studies, especially those by Ekstrand et al. [26] are applicable to the book domain we have chosen. If our results are comparable, this would be another step toward formulating generally applicable statements about user preferences in recommendation systems, as the results have been confirmed in two different domains. We formulated the following research objectives to contribute to this development.

1. **Understand User Preferences:** The primary objective of this research is to have a thorough understanding of consumer preferences regarding book recommender systems. This means investigating the fundamental aspects that drive users' perception of different recommender algorithms. By doing so, we hope to shed light on the

complexities of user decision-making in the context of book suggestions.

2. **Assess User Satisfaction:** Another objective is to evaluate user satisfaction with three commonly used collaborative filtering recommender systems. We want to know how well these systems meet users' expectations and give recommendations that are relevant to their interests.

3. **Analyze User Behavior and Interactions:** This briefly treated objective focuses on examining user behaviour and interactions with book recommendation systems. We hope to gain useful insights into how users interact with these systems and how to increase the effectiveness of the design of a similar study in the future.

In order to achieve these objectives, the following research questions (RQs) have been formulated:

- **RQ1:** What are the primary factors that influence user preferences when it comes to the first impression of book recommendations?
- **RQ2:** What differences do users perceive between the lists of recommendations?
- **RQ3:** Do different user groups hold varying preferences for recommender algorithms?
- **RQ4:** How do objective algorithm performance metrics relate to users' subjective assessment of recommender algorithms?
- **RQ5:** What are common patterns and behaviours in how users interact with book recommender systems?

The scope of the study is limited to collaborative filtering algorithms only. The study examined an Item-Item, User-User, and ALS algorithm, all of which are widely used. These algorithms have been previously addressed in related research [26], facilitating easier comparability. Additionally, these algorithms are relatively easy to implement. This allowed for a greater focus on analyzing user preferences rather than intricacies in the implementation. Consequently, the research does not provide insights into how user preferences might be for content-based or neural network collaborative filtering algorithms. Furthermore, our investigation measures the user perception of each participant once and does not consider changes in perception over a specific period, acknowledging the fact, that perceptions may be influenced by factors that vary across different timeframes.

This thesis is structured into several chapters to provide a logical framework. In section 2 a literature review is conducted and will present relevant literature on recommender systems, user perception, and other related topics. This will provide the theoretical foundation for this study. Following in section 3, the methodology will be explained. This will detail research methods, data collection techniques, and data analysis procedures. In section 4 a clear and unbiased presentation of the research results will be conducted. Section 5 is dedicated to interpreting the results and answering the research questions based on the research results. The final section serves as a conclusion, summarising the main findings and exploring potential avenues for future research.

Chapter 2

Related Work

In this section, related work and several key terms and concepts, as well as important papers will be mentioned and explained. The following topics will be addressed: Recommender systems in general, book recommender systems, user perception of recommender systems and human-centred research, and bias in recommender systems.

2.1 Recommender Systems

Recommender systems (RS) operate based on the analysis of user data, item characteristics, and historical interactions to generate personalized recommendations [13]. The core idea is to find patterns or similarities among users and items and leverage these patterns to make predictions [2]. The techniques used in recommender systems encompass a variety of artificial intelligence methods, including machine learning and data mining [100, 68]. These techniques are applied to two types of data. User data, which includes user preferences, demographic data, and historical data, such as purchase histories, ratings or implicit interactions with an interface or items [56]. Item Data, the second type, can consist of explaining features such as colour, dimension, or other attributes of an item. Such data is considered to assess the characteristics of items in the recommendation process.

Recommender systems can be categorized into several types, each with distinct approaches and characteristics: **Content-Based Recommender Systems** (CB): These systems recommend items based on their qualities and content. They are well-suited for proposing things with similar features to the interests demonstrated by the user. There has been extensive research on these types of algorithms in the early days of recommender systems, which was summarised in [66, 55]. The advantages of content-based filtering are user independence when building the recommendations since the recommendations are specific to this user, easier provision of explanations and transparency of the recommendations, and the possibility to capture the specific interests of a user resulting in recommending niche items. Limitations are among other things the proneness to overspecialization, difficulty in generating descriptive data for items in certain areas, and the great need for domain knowledge [6].

More relevant for this research are **Collaborative-Filtering Algorithms** (CF). These types of algorithms rely on past user-item interactions. Frequently used and well-researched methods include: User-based collaborative filtering recommends items to a user based on the preferences of other users with similar tastes [40, 34]. Item-based collaborative filtering on the other hand suggests items similar to those a user has interacted with. Algorithms calculate item similarity to find the most similar item based on user interactions. Used similarity metrics include cosine-based approaches [20, 53], correlation-based solutions using the Pearson coefficient [80], or more expressive solutions which analyse nonlinear and higher-order relationships among items [98]. Matrix factorisation techniques decompose the user-item interaction matrix into latent factors. There has been substantial research on how to handle the resulting sparse item-user matrix resulting in many different algorithms. Solutions include Non-Negative [57], Bayesian probabilistic [78], Factorisation machines [75], Singular Value Decomposition (SVD) [28, 9], the SVD++ algorithm which, as an extension of the previous SVD, takes into account implicit feedback [47, 37], and Alternating least square (ALS) [106, 89] approaches. The algorithms selected for the later coming analysis of user perception are ALS [106], the user-based CF from [34], and the item-based solution introduced in [20]. The advantages of CF techniques include no dependence on domain knowledge and capturing user interests over time as well as inherent subtle characteristics. Key disadvantages are the user cold-start problem, which states that the system cannot recommend items to new users who have not interacted with the system yet and the item cold-start problem, where the system cannot recommend an item that users never selected before, since it is not in the user-item matrix [52].

To get the best of both worlds, past work combined the strength of both content-based and collaborative filtering to create **Hybrid Recommender Systems**. As promising the approach to overcome the limitations of individual recommendation methods sounds, challenges such as integration and computational complexity should be carefully addressed. Additionally a study in 2017 revealed [14], the main challenges are still data sparsity [36, 95, 29] and the cold start problem [58, 18, 7]. The main advantages are as mentioned before the increased accuracy of the recommendation, more diverse recommendations due to the usage of both techniques and increased personalisation by considering both the user preference and the item characteristics in the recommendation. The disadvantages are the increased complexity making the integration more compound, the hard-to-overcome cold start problem, and the dependence on a substantial amount of data, which can result in problems when data is not of high quality or worse not available [12].

Advancements in deep learning have led to the development of **Neural Network-Based Collaborative Filtering Recommender Systems** (NCF). These systems utilise neural networks to model complex-user-item interactions and have shown promising results in improving recommendation quality, especially in the diversity of recommendations [5]. [33] introduces a neural network-based collaborative filtering model that combines matrix factorisation and neural networks for improved recommendation accuracy. In [16] the Wide & Deep model, which combines the strengths of linear models for memorisation and deep learning for generalisation in recommendation tasks, is presented. The use of autoencoders, a type of neural network architecture, for collaborative filtering tasks is explored in [82]. It demonstrates the effectiveness of neural collaborative filtering with explicit matrix factorisation. Neural network-based recommender systems can outperform state-of-the-art results in collaborative filtering concerning RMSE, ease the cold start problem, and are scalable and robust to deal with large-size datasets [87]. Due to the behaviour as a black box, providing explainable predictions can be a really challenging task. Also, deep learning is known to be data-hungry, which can be a problem if there is not much data present in the employed domain [101]. Another downside is the need for extensive hyperparameter tuning [76].

2.2 Book Recommender Systems

In the realm of personalised content recommendations, book recommender systems play an important role in enhancing user engagement and satisfaction [92]. Researchers in [21] used user-based collaborative filtering and cosine-similarity to generate recommendations. Pearson correlation was also discussed however, it is not clear how this similarity measure was used in the final algorithm. The cold start problem for new users was addressed by providing them with a list of the current top ten books with the highest average rating. Users are required to rate these books, thereby generating data for the collaborative filtering algorithm. To overcome the same problem for new books, a newly added book was given the average rating of books by the same author already present in the data set.

In [59] a hybrid recommendation system is presented which uses content-based and collaborative filtering methods to create recommendations. The algorithm used not only ratings given by the existing users for the recommendations but also investigated associations and relationships among a large dataset of books using association rule mining. The researchers implemented the proposed algorithm in a new website application.

A hybrid approach was also used by [92] where researchers aimed to overcome the problem of selecting appropriate books from a large set of candidate books in a library scenario. They used cosine similarity and a k-means algorithm. To handle the problem of data sparsity user clustering is performed, resulting in 15 user-category matrices. Further analysis showed, that the proposed algorithm outperformed its' single algorithms in matters of precision.

Digital libraries were also investigated by [54]. They used deep belief network algorithms to construct digital library scenarios based on the construction of a personalised recommendation system for library resources. An ontology similarity is calculated and used as a basis for a personalised recommendation algorithm.

In [90] ontology-based recommendation was also explored in the context of editorial products. A semantic recommender system developed in collaboration with Springer Nature was presented, which should support the user in selecting relevant publications for research venues. Users are also presented with a graph view, which explains why a certain product was recommended and how its associated topics intersect with the ones characterising the input conference.

The usage of deep learning in book recommendation was explored in [70]. The researchers proposed a hybrid approach combining the strength in accuracy of collaborative filtering algorithms and addressing its problems by adding deep learning. The resulting algorithm uses embeddings for representing users and items to learn non-linear latent factors. Also, the cold start problem is handled by integrating side information about users and items into a very deep neural network. Results of testing the algorithm on the Book-Crossing dataset showed, that it is capable of outperforming existing methods in both non-cold start and cold start cases.

Goodreads, a social website for 'readers and book recommendations', has developed a large user base for book-based social web services [91]. In a blog post in 2011, they introduced personalised recommendations for the user of the web service [19]. The exact method implemented is not stated in the post, yet the given description of the algorithm suggests the implementation of an item-based collaborative filtering algorithm working on the bookshelves a user can create. Other recommendation methods include the popular 'reader also enjoyed' [11] category when inspecting the goodreads page of a book, which recommends books that are on the same shelves created by a user who read the current book. After these implementations, there have been no innovations regarding further recommendations, which could be linked to the Amazon takeover in March 2013.

Speaking of Amazon, the complex online shopping platform also offers recommendations, when inspecting a book. [51] investigated the user strategies that emerged in those environments and found out by observing and interviewing internet shoppers, that after a keyword search, recommender systems played an important role in helping users find the books they wanted. Amazon, equally like Goodreads, presents a ‘related products‘ category when browsing items. Unlike the algorithm of Goodreads though, Amazon uses an item-based approach for recommendation [53, 32] and can be commercially motivated since authors can pay for the placement in the related products list.

2.3 Human-centred Research

In the rapidly evolving landscape of technology and information systems, the significance of understanding and prioritising the human element cannot be overstated. Human-centred research (HCR) and Human-Computer Interaction (HCI) serve as disciplines at the intersection of technology and user experience. HCI is a multidisciplinary field, that combines principles from computer science, psychology, design, and sociology to investigate how humans interact with computers and computational systems [50].

HCI plays an important role in recommender systems which is summarised by conducting an overview of the history of recommender system research from a human-centred perspective by [46]. The authors showed, how looking at interface and algorithm studies has advanced our understanding of how system design can be tailored to users’ objectives and needs. The influence of external factors has also been investigated. Several examples of next-generation approaches to further deepen existing knowledge, which emerged in recent years, were given:

In multiple papers, Zhao et al. investigated how to incorporate non-selection by a user as input. Starting by analysing eye tracking [103] combines the resulting data with browsing data to gather gaze patterns. Results show, that by incorporating this data in recommendation, the accuracy was boosted significantly. Later in 2018, a rich model for interpreting users’ failure to select a recommended item was built based on previous research [104]. In 2019, Zhao et al. incorporated this non-selection data into eight different single models and three associated hybrid models on a user browsing data set. They also designed a novel model based on recurrent neural networks. They report, that this new model outperforms others in several metrics [102].

Another promising avenue was gaining a deeper understanding of choice overload. Choice overload describes the phenomenon, that users have more difficulty choosing from a large list and as a result are less satisfied with their choices [17]. Researchers in [96] overcome this problem by investigating, whether the diversity of items in a recommended list affected the effort to make a choice. They found out, that not only did more diverse lists remove effort to make a choice, but diversity also led to higher user satisfaction. The authors also encouraged other researchers to pay attention to choice overload and to create shorter more diverse recommendation lists.

For this research especially important is the analysis of user perception of algorithmic differences. In 2014 Ektrand et al. [26] conducted a large-scale user study on the differences in user perception of different recommender algorithms. The goal was to understand how users perceive the performances of algorithms, against objective measures, concerning accuracy, diversity, novelty, personalisation, and overall satisfaction. The analysis of the user data showed the surprising result that excessive novelty was harmful to user satisfaction. In contrast to novelty, diversity appeared to have a positive influence, which is intriguing, since increasing diversity often leads to a decrease in accuracy [99, 105, 107]. This result coincides with the findings from choice overload research, as well as poses the question of whether the past focus on accuracy may have been harmful to user satisfaction [61, 43]. To further research this hypothesis, a field study in 2015 followed by Ekstrand et al. [25] let users choose their own algorithm on the MovieLens website [63]. Results showed if the user had the choice to pick one of the three algorithms from the previous research, they actually picked the one identified by the model.

A fundamental part of this research will be trying to replicate the results of Ekstrand et al. by using a similar approach which will be explained in the next section. By expanding the field of research from movie recommendation to book recommendation, it can be investigated whether the findings were domain-dependent or are also congruous in the realm of books. This result will deepen the knowledge of user perception of recommendation algorithms.

2.4 Bias in Recommender Systems

Another crucial part of this thesis will be investigating potential biases in recommender systems. However, due to later explained design choices the algorithm per se will not be considered. Rather potential biases in different user groups will be the subject of investigation. Understanding the impact of bias in recommendation systems is paramount as it directly influences user experiences and decision-making. The work of Ekstrand et al. in 2018 [24] conducted research based on advances in information retrieval evaluation which showed, the importance of considering the distribution of effectiveness across diverse groups of varying sizes. They used offline evaluation and a utility-based metric of recommendation effectiveness to explore whether different user demographic groups experience similar recommendation accuracy. The authors found demographic differences in measured recommender effectiveness across two data sets containing different types of feedback in different domains.

Researchers in [65] conducted a 'fish-eye view' of the landscape of the sources of bias, and the solutions proposed to address them from a broad cross-domain perspective. Recommender systems were one of four areas of research. They showed steps towards comprehensive treatment bias detection, fairness management, and explainability management.

There has been versatile research on how algorithms can be biased towards distinct user groups in different domains. [38] for example, investigated bias in job ads on LinkedIn and Facebook. They published the same job ads on these websites, where only the demographic data of the applicants differed. The results showed a skew by gender in ad delivery on Facebook. At LinkedIn, the researchers could not find skew in ad delivery. A similar study was conducted by [84] investigating Facebook's online targeting of advertisements. The researchers showed, that disallowing advertisers to use sensitive attributes, such as ethnic affinity or gender, when targeting ads related to housing or employment or financial services, is not enough to diminish the problem of discrimination in online advertisement. The results of systematically investigating the different targeting methods provided by Facebook showed, that malicious advertisers can create highly discriminatory ads without using sensitive attributes. These results are in line with those by [3]. The study found, that both the advertiser's budget and the content of the ad each significantly contribute to the skew of Facebook's ad delivery. They critically observed a significant skew along gender and racial lines for 'real' ads for employment and housing opportunities. A study by [88] showed, that Google's ad delivery algorithms were also not free from bias.

A study in 2017 by Edelman et al. [22] revealed, that other domains were also affected by racial discrimination. In an experiment on Airbnb, an online marketplace for short- and long-term homestays and experiences, the researchers found that applications from guests with distinctively African American names are 16% less likely to be accepted relative to identical guests with distinctively white names.

In [27] user behaviour towards awareness of biased algorithms has been researched. By using a cross-platform audit technique, they analysed online ratings of 803 hotels across three hotel rating platforms. Results showed, that Booking.com's algorithm biased ratings, particularly low-to-medium quality hotels, significantly higher than other sites. By analysing reviews of 162 users who independently discovered the mentioned bias, the authors aimed to understand if, how, and in what ways users perceive and manage this bias. Users tried to raise awareness of the bias by reverse-engineering the rating algorithm, making efforts to correct the bias, and stating that their trust was broken.

Chapter 3

Methodology

The research design employed in this study is characterized by a quantitative approach, characterized by the systematic collection and analysis of data to find patterns, correlations, and statistical significance [42]. This methodological framework enables us to carefully evaluate and quantify the variables under research using various statistical techniques such as surveys, experiments, or structured observations. The section is divided into sections, each of which describes a different aspect of the process. The components are either required to conduct the study or are required to answer the research questions.

3.1 Experiment Design

We needed to consider the optimal data collection methods and research methodologies to address our defined research questions. Quantitative research is a deductive strategy that seeks to establish links and patterns through systematic data collection methods such as surveys, experiments, or content analysis. It uses numerical data and statistical analysis to measure and quantify effects. Qualitative research, on the other hand, takes a more interpretative and inductive approach, focused on understanding the context, meanings, and experiences related to a certain phenomenon. Interviews, focus groups, and content analysis are common qualitative methodologies that allow for a more in-depth investigation of subjective perspectives. Since we want to investigate patterns in our data on how users perceive recommender algorithms and which factors influence the users' choice, we chose a quantitative approach by conducting a study to collect user data with a survey. We created an experiment in which users were presented with lists of book recommendations, each containing ten items, generated by different algorithms based on explicit user ratings. The participants were also asked to fill out a questionnaire to assess the perceived differences between the lists and various factors. In the following section, we will explain and describe the created system used by the participants to collect the study data.

3.2 System Design

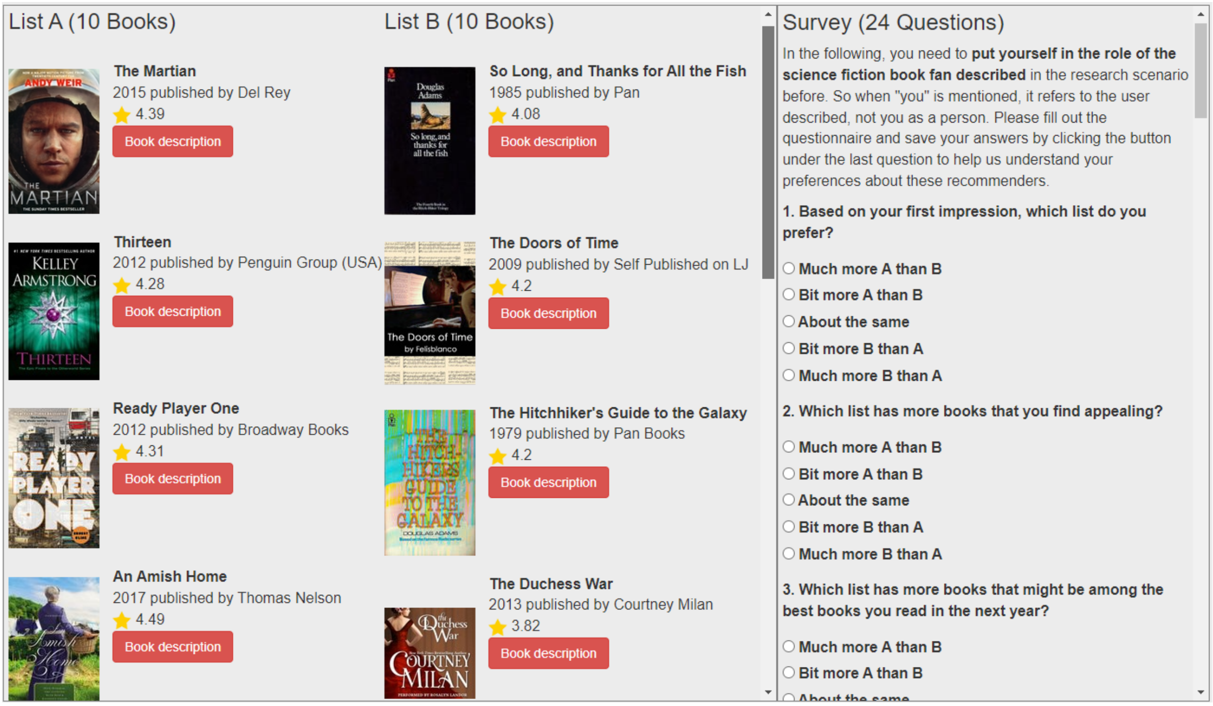


Figure 3.1: Screenshot of the user interface of the experiment. Buttons and covers are clickable to access more details. Both sides are scrollable to reveal the remaining content.

Figure 3.1 shows a screenshot of the study interface the user was presented with. The participant has immediate access to information on the design of the cover, title, publishing date, publisher, and average rating of each book. To become more familiar with the books, the user can also click the ‘Book description‘ button, which displays a summary of the content, or click on the cover, which opens the associated Goodreads [31] page in a new tab. As this experiment was not conducted within an existing platform, as Ekstrand et al. [26] did with MovieLens [63], existing historical user data could not be utilized. This limitation implies that it would be very challenging to generate personalized recommendations for each participant in the study. Therefore, a research scenario was constructed, introducing a science fiction fan with previously read and rated books. Participants were instructed to conduct the study ‘through the eyes‘ of the introduced fan. This approach allowed us to generate at least personalized book recommendations for the constructed user, as further explained in section 3.5. The complete scenario is depicted in Figure 3.2.

Research Scenario

Imagine you are a big fan of science-fiction books and are currently looking for a new book to read. You've already read *Frankenstein*, *Foundation*, *The Stars my Destination*, *Solaris*, *Dune*, *The Moon is a Harsh Mistress*, *Hyperion*, *Jurassic Park* and other similar books. You gave them each a good rating (3-5 stars). To assist you, you are using the help of different recommendation systems, each of which has created a top-10 list for you. The first book in each list represents the system's best-fit recommendation. To access more details, simply click on the cover page of each book, which will open the respective GoodReads webpage in a new tab. Please familiarize yourself with each book on the lists and complete the questionnaire afterwards.

Figure 3.2: Research scenario. The cursive book titles were clickable and opened the corresponding Goodreads page in a new tab.

When comparing our approach to those of previous research, such as Ekstrand et al., our study has natural limitations due to the research scenario. In contrast, previous research had the advantage of accessing historical user data, as these experiments were conducted on existing recommendation platforms, here on MovieLens [63]. The study on movie recommendation systems had a total of 582 participants, all of whom were active users of the website and had already provided at least 15 ratings [26]. It can be assumed that the collected data is of high quality since, on the one hand, personalised recommendations could be generated for each participant, and on the other hand, users already had experience with the system and inherently possessed a certain interest in movies. The quality of our response data depends on how well each participant is familiar with the science fiction topic or how much time was invested in becoming acquainted with potentially unfamiliar books. Additionally, it is not guaranteed, that the participants of our study necessarily have an interest in book recommendations or reading itself, but we tried to minimise this effect by recruiting users who could be interested in these topics. More on the user recruitment in section 3.3. Despite these limitations in aspect to previous research, personal preferences can be measured through the responses in any case.

After the user had successfully participated in the study and provided responses to individual questions, they were given the opportunity to participate in the study again and evaluate a different combination of lists. After each successful submission, the displayed content was altered (more details in section 3.5). This could be initiated by clicking a 'Back to

Survey‘ button.

To measure how often and with which parts of the user the interactions occurred during the experiment, specific actions were logged. Log data was primarily collected during user interaction with the experiment interface using a JavaScript logger. The main purpose of the data was to analyse user behaviour to gain valuable insights into which design decisions and components were well-received or poorly received. This will help similar experiments in the future to help make a choice on certain design decisions and to further improve the interface. Additionally for this experiment, it serves as an additional quality control measure for submissions to exclude them when there was little or no interaction with the interface, making it impossible to draw qualitative conclusions.

Two aspects were logged: scrolling behaviour in the list area of the interface (see the left part of Figure 3.1) and the click behaviour on individual buttons and cover pages. Each button and cover has a unique ID for later analysis to determine which book received the most clicks for more information. A log message consists of a timestamp, the participant's session ID, and a description of the user interaction. A typical message for a button click looks as follows:

Session ID: bg621uij Time: 2023-11-27T12:31:58.613Z - Button clicked: Stoner Description

We expect that the user interaction with buttons will be greater than with the covers and that familiar books will tend to have fewer clicks than unfamiliar ones.

3.3 Users and Context

As mentioned above, since the study operated without a known platform, it was hosted on an unknown website. Participants also had to be recruited without relying on existing users. To maintain the quality of responses, efforts were made to recruit participants with relevant interests. This included individuals who either generally consume a lot of books or have a significant interest or expertise in the science fiction genre. Recruitment sources included subreddits such as r/books [72], r/scifi [74], r/printsf [73], as well as blogs like ‘Der Buechernarr‘ [49], a German-language literature blog.

3.4 Dataset

To generate recommendations for the user created in the scenario, book reviews are necessary. Two datasets from the project [93, 94] conducted at the end of the year 2017, scraping data from Goodreads, were chosen. In this project, all publicly accessible bookshelves on the Goodreads platform were scraped. These bookshelves can be personally created by any user and represent a way to organize books that have been already read or saved for future reading. User IDs were anonymized.

To train the algorithms and generate book recommendations, a dataset of book reviews was essential. This dataset includes approximately 15m multilingual reviews of around 2m books and 465k users. To generate more accurate recommendations, only reviews from users who have published at least 15 reviews were considered. While this restriction reduces around 70% of users, it only decreases the number of reviews by approximately 9%. Therefore, a significant number of reviews still remain, justifying this limitation.

To provide the user with the information for each book, mentioned in section 3.2, another dataset consisting of information about the reviewed books was necessary. This dataset is approximately 2 gigabytes in size and covers the same number of books as the other dataset used. The attribute 'book_id' is consistent across both datasets, allowing for a straightforward connection.

3.5 Algorithms

To create the user outlined in the research scenario, a specific dictionary was created, which later can be used to recommend items. The user was assigned the 'user_id' -1 to ensure uniqueness. The selected books are considered good or among the best in public opinion in their specific genres. The ratings for these books are consistently good (3-5 stars), indicating that the user 'liked' all the selected books. The entire reading history can be seen here:

Listing 3.1: User configuration

```
# Creating the research scenario user
# key: value -> book_id: rating
research_user = {
    18490: 4, # Frankenstein
    29579: 4, # Foundation
    333867: 4, # The Stars My Destination
    95558: 3, # Solaris
    234225: 5, # Dune
    16690: 5, # The Moon is a Harsh Mistress
    77566: 5, # Hyperion
    7677: 5, # Jurassic Park
    5470: 4, # 1984
```

```
    5129: 3, # Brave New World
    4981: 4, # Slaughterhouse-Five
    2767052: 5, # The Hunger Games
    830: 4, # Snow Crash
    7613: 3, # Animal Farm
    227463: 4 # A Clockwork Orange
}
```

For this study, three widely-used collaborative filtering algorithms were tested, all implemented with the LensKit [23] package and also used in Ekstrand et al.'s experiment, with the exception of the Alternating Least Squares (ALS) algorithm used instead of the FunkSVD algorithm. This decision was made based on a hint in the LensKit documentation, stating that ALS-based algorithms are less sensitive to hyperparameters. The default settings proposed by LensKit for each algorithm served as the starting point for parameter tuning. After manual investigation of the results, the configurations used by Ekstrand et al. [26] in their experiment were applied.

Since two of the three algorithms, with these settings, found too few suitable neighbours and therefore did not generate a complete Top-10 list, the following configurations ended up with:

- Item-Item CF [20] with 20 neighbours, cosine similarity, item mean centring, a neighbour threshold of 0.1, and requiring 1 neighbour to make predictions (Listing 3.2).
- User-User CF [34] with 30 neighbours, cosine vector similarity between users, a neighbour threshold of 0.1, and requiring 1 neighbour to make predictions (Listing 3.3).
- ALS CF [106] with 50 features, 20 iterations per epoch, regularisation factor of 0.1, and damping of 5 (Listing 3.4).

With these three configured algorithms, a Top-10 list was generated for the created user by each algorithm. The complete recommendation lists generated by each algorithm can be viewed in Appendix A.

The participant in the study was then presented with two out of the three lists in the experiment. This decision was made for ease of comparison (it's simpler to compare two lists than three) and to quantify differences between them using a questionnaire. To evenly distribute the statistical presentation of the lists, a solution similar to a Latin square rotation [42] was implemented. Each list was paired with every other list, taking on the role of List A and List B alternately, to eliminate a natural preference for List A. In the end, there were six unique combinations of lists, and they were rotated through one by one after each successful submission of a participant's survey responses.

We did not want to employ a between-subject design in which each participant is only shown one list at a time since we wanted to investigate the differences between the user perception of these recommendation algorithms. If the participants were presented with only one recommendation at a time, they would lack the context of other lists, which could be problematic. Absolute judgments can be considerably more challenging than making relative judgments in this context [26, 35].

Listing 3.2: Item-Item Configuration

```
nnbrs = 20
min_nbrs = 1
min_sim = 0.1
feedback = 'explicit'
center = True

algo_ii = item_knn.ItemItem(nnbrs=nnbrs, min_nbrs=min_nbrs, min_sim=min_sim,
    feedback=feedback, center=center)
```

Listing 3.3: User-User Configuration

```
nnbrs = 30
min_nbrs = 1
min_sim = 0.1
feedback = 'explicit'
center = True

algo_uu = user_knn.UserUser(nnbrs=nnbrs, min_nbrs=min_nbrs, min_sim=min_sim,
    feedback=feedback, center=center)
```

Listing 3.4: ALS Configuration

```
features = 50
iterations = 20
reg = 0.1
damping = 5

algo_als = als.BiasedMF(features=features, iterations=iterations, reg=reg,
    damping=damping)
```

3.6 User Survey

The decisions made in compiling the user survey were significantly influenced and guided by past research by Ekstrand et al. [26]. On the one hand, this aimed to achieve a form of comparability, and on the other hand, Ekstrand's work demonstrated that the methods used were highly effective for their goal. The explanations in this chapter and the content of the study's questionnaire are thus strongly inspired by this research.

At the beginning of the questionnaire, participants are reminded to assume the role of the person described in the scenario when answering the questions (see the upper right figure). Therefore, when the question addresses "you," it refers to the depicted user in the scenario, not the respective participant.

The first question explores which of the two lists the participant would prefer based on their initial impression. There are five response options ranging from 'Much more A than B,' 'Bit more A than B,' 'About the same,' 'Bit more B than A,' to 'Much more B than A.' This pattern continues for the next 22 questions, which aim to capture the personal perception of various aspects of the recommendation lists. The following five factors will be investigated [26]:

- **Acc** Accuracy - the recommender's ability to find 'good' books.
- **Div** Diversity - the diversity of the recommended items.
- **Und** Perceived personalisation ('Understands Me') - the user's perception that the recommender understands their tastes and can effectively adapt to them.
- **Sat** Satisfaction - the user's overall satisfaction with the recommender and their perception of its usefulness.
- **Nov** Novelty - the propensity of the recommender to suggest items with which the user is unfamiliar.

For each factor, 4-5 questions should be selected. A successful factor analysis requires 3 questions per factor. The additional questions allow us to later eliminate questions that do not contribute much information. The factor analysis will measure how well each question assesses its intended factor, and weak, low-loading questions will be removed.

The selection of questions for each factor was decided, as described above, based on Ekstrand's question set. This research, in turn, relies on Knijnenburg et al. [45], who developed a complete questionnaire with factors. We adapted the questions to the context of books to make the formulations meaningful. Table 4.2 shows all the questions.

We expect similar results to Ekstrand et al.'s research: the first impression is influenced by perceived satisfaction and novelty. Which, in turn, has an impact on diversity and satisfaction.

3.7 Demographic Data

In the first part of the study, participants were asked for demographic data, which was anonymously stored and used exclusively for research purposes. The goal was to gather both personal and 'expert data.' The latter indicates how well participants are acquainted with books, especially in the science fiction genre, and what preferences they have regarding reading. The following data were collected:

- Age: Age is categorized into seven age groups.
- Gender
- Book Format: Participants are inquired about their preferred reading format. Options include physical books, e-books, audiobooks, online articles, and others.
- Book frequency: Participants are asked to estimate how many books they typically read in a year. Data from [86] revealed, that the average U.S. resident read 12.6 books or roughly 1 book per month in 2021. So we chose this to be our centre point and created five categories around it: 0-4, 5-9, 10-14, 15-20, and more than 20.
- Preferred genres: Participants are presented with a selection of the most popular genres to choose from. In case a genre is missing, users have the option to select "Other Fiction" or "Other Non-Fiction" to include any additional genres not listed. Again this data gains insights into the reading habits and preferences of the particular participant. Furthermore, contributions from a participant who has chosen Science Fiction or other related genres are particularly interesting for the research, as the user exhibits greater expertise.

This data was collected to extend the Latent Factor Analysis from section 3.6 with additional factors. The aim is to investigate how personal characteristics such as gender and age can influence various factors, as well as the extent to which domain knowledge can have an impact. According to [45], age, gender, and domain knowledge influence users' perceptions, experiences, and behaviours. In these experiments, age had a positive impact on the interaction with the interface, which in turn had a positive effect on the perceived effectiveness of the system. On the other hand, gender had a negative impact on the perceived accuracy of recommendations. Finally, expertise or domain knowledge showed a positive influence on both the perceived accuracy and diversity of recommendations, as well as a positive impact on satisfaction. We expect similar results in the context of book recommendations.

3.8 Objective Metrics

To expand the analysis capabilities, objective metrics were calculated in addition to participants' subjective perceptions of the recommendations. These metrics investigate the selected algorithms based on Accuracy, Novelty, and Diversity.

To determine the accuracy of an algorithm, the dataset was initially split into training and testing partitions, with the test partition containing the last 5 ratings of each user. It's important to note that these do not necessarily have to be the chronologically last 5 ratings. The algorithm was then used to estimate the ratings in the test partition. During this process, the overall Root-Mean-Squared Error (RMSE) was calculated, serving as the accuracy metric.

To measure novelty, the mean popularity rank of the recommended book list was calculated. The popularity of a book was firstly determined by counting how often the specific 'book_id' appeared in the entire dataset — essentially, how many reviews were written for that book. Since the dataset includes different 'book_ids' representing varying editions, prints, and versions but effectively the same book, instead of the id the exact titles were counted. Using these popularity scores, a ranking was created, giving each unique title a popularity rank. A high average popularity rank indicates that the list includes more obscure books.

Diversity was computed using intra-list diversity, employing cosine similarity between tag genome vectors as the itemwise similarity function [26]. The tag genome is a structure containing scores indicating the degree to which tags apply to items [48]. We used the dataset, by [48] to make these computations. This dataset was created using data from the collection we also used for this project, containing scores of 727 unique tags for 9,374 unique books. Since the book recommendation lists partly include obscure and unknown books, and the dataset unfortunately only includes a selected portion of all books, only half of the books in a list could be considered when calculating intra-list similarity. Therefore, this metric should be interpreted with caution, as it can at best reveal a tendency. A high score indicates little diversity within the list.

Chapter 4

Results

Over a period of one month, 150 individuals started the study, with 112 of them successfully completing it. To adhere to ethical guidelines, submissions indicating the age group "under 18" were removed. Additionally, data quality checks were conducted. Submissions were filtered out if they incorrectly answered the attention check question or if no scrolling data were available. After these adjustments, 97 qualitative submissions remained.

As no statistical differences were observed in the conditions where the same two algorithms appeared in reversed order, the subsequent analysis combines each pair of conditions. The conditions I-I vs. ALS and ALS vs. I-I are combined into I-I vs. ALS, and so on. This makes further analyses more straightforward.

Based on the initial impressions, users generally found ALS better than Item-Item ($p <$ 0.001) and User-User ($p < 0.000006$). These values are also reflected in Table 4.1. When users were presented with ALS in a condition, they preferred the initial impression of ALS (choosing 'Bit more' or 'Much more' towards ALS) with an 88.7% probability. It is less clear in the I-I vs. U-U condition. While the initial impression of User-User is preferred over Item-Item ($p < 0.0004$) *when* a preference is stated, a preference is only stated in 67% of cases. The other 31% of the users did not express an initial preference between the two lists. Table 4.2 summarizes the responses to all our study questions based on the algorithm constellation.

Cond. (*A* **v.** *B*)	*N* **Total**	*N* **Tendency**	**Tend.** *A*	**Tend.** *B*	**% Tend.** *B*	*p*
I-I v. ALS	27	23	5	18	66.6%	0.001
I-I v. U-U	35	24	5	19	54.3%	0.000
ALS v. U-U	35	32	26	6	17.1%	0.000

Table 4.1: First impression of the algorithm by condition. p-values are for two-sided proportion tests, H_0: $\frac{a}{b} = 0.5$

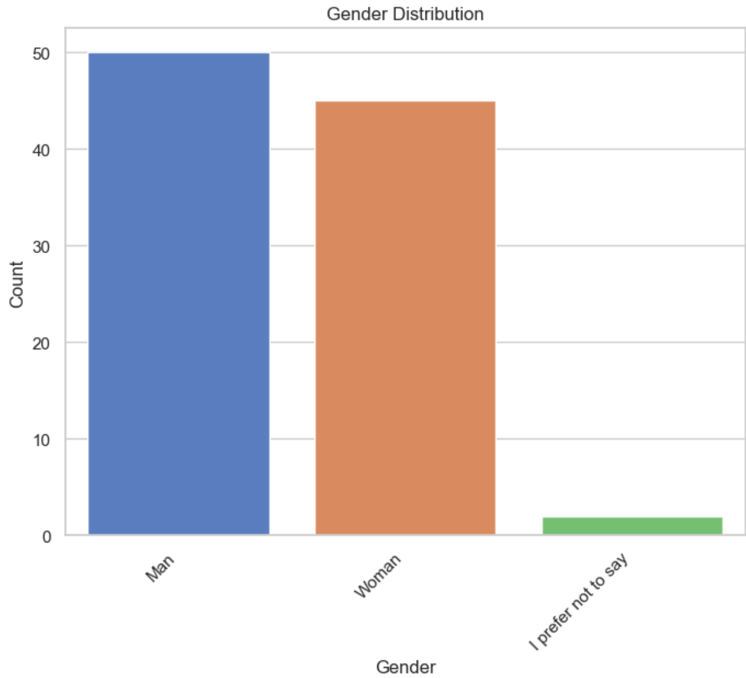
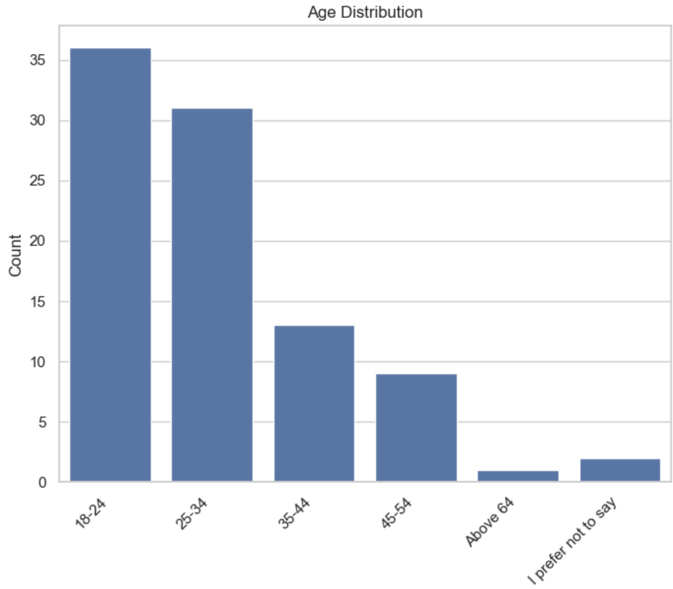


Figure 4.1: Net gender distribution of the experiment participants.

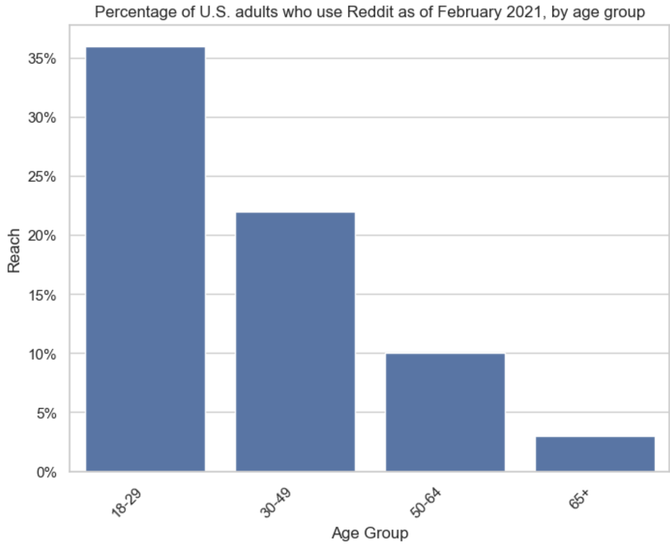
4.1 Demographic Data

Figure 4.1 depicts the gender distribution of the participants in the experiment. Approximately 52% indicated their gender as ‘Man‘, and about 47% as ‘Woman‘. The remaining percentage chose not to disclose their gender. The distribution between ‘Man‘ and ‘Woman‘ is roughly balanced. Figure 4.2a illustrates the distribution of age groups among the participants in the experiment. Notably, the ‘18-24‘ and ‘25-34‘ age groups stand out, together representing approximately 70% of the total participant pool. This can be explained by the platforms used for user recruitment (see section 3.3. Reddit exhibits a similar age distribution with the highest share being ‘18-29‘, as depicted in Figure 4.2b.

Participants were asked about their preferred way of consuming books, with multiple choices allowed. Figure 4.5a displays the distribution in response to this question. Almost all participants mentioned the physical book as a preferred medium. Following with about half the mentions is the e-book. Online articles and audiobooks come next and are roughly equal in mentions. The participants were also asked about the average number of books they read per year. Figure 4.5b displays the results of this question. The assumption from section 3.7 that the average U.S. citizen reads 12.6 books per year is not confirmed by the experiment. In this case, the average is slightly lower, at approximately 8.2 books per year.



(a) Net age group distribution of the experiment.



(b) Percentage age group distribution of U.S. Reddit users.

Figure 4.2: Comparison of the age group distribution of the experiment with that of Reddit. Source graph on the right: [85].

Most participants indicated the categories ‘0-4‘ and ‘5-9‘. Following that were the responses given for ‘More than 20‘.

To determine the most popular genres among the participants, all responses for each genre were accumulated. The results are presented in Figure 4.3. Not surprisingly, the most popular genres are ’Mystery, Thriller, and Crime’ and ’Fantasy.’ These genres are also widely popular and commonly read in general. The third-place ranking of Science Fiction is encouraging, as it indicates that 47% of participants have at least a basic understanding of the domain the scenario sketched, promising qualitative insights. To further analyze the genre distribution, genres were examined based on gender. Genre occurrences were divided by gender. To keep the resulting graph clear, the integration of the two responses where gender was not specified, was excluded. The occurrences for each genre per gender were normalized by dividing the sum by the total number of all genre responses for that gender to compensate for the statistical imbalance between men and women in the dataset. The resulting grouped bar chart is shown in Figure 4.4. Notable are the genres ’History,’ ’Political,’ and ’Romance,’ with the first two being more preferred by men and the latter more by women. There were no significant differences in the genre of science fiction.

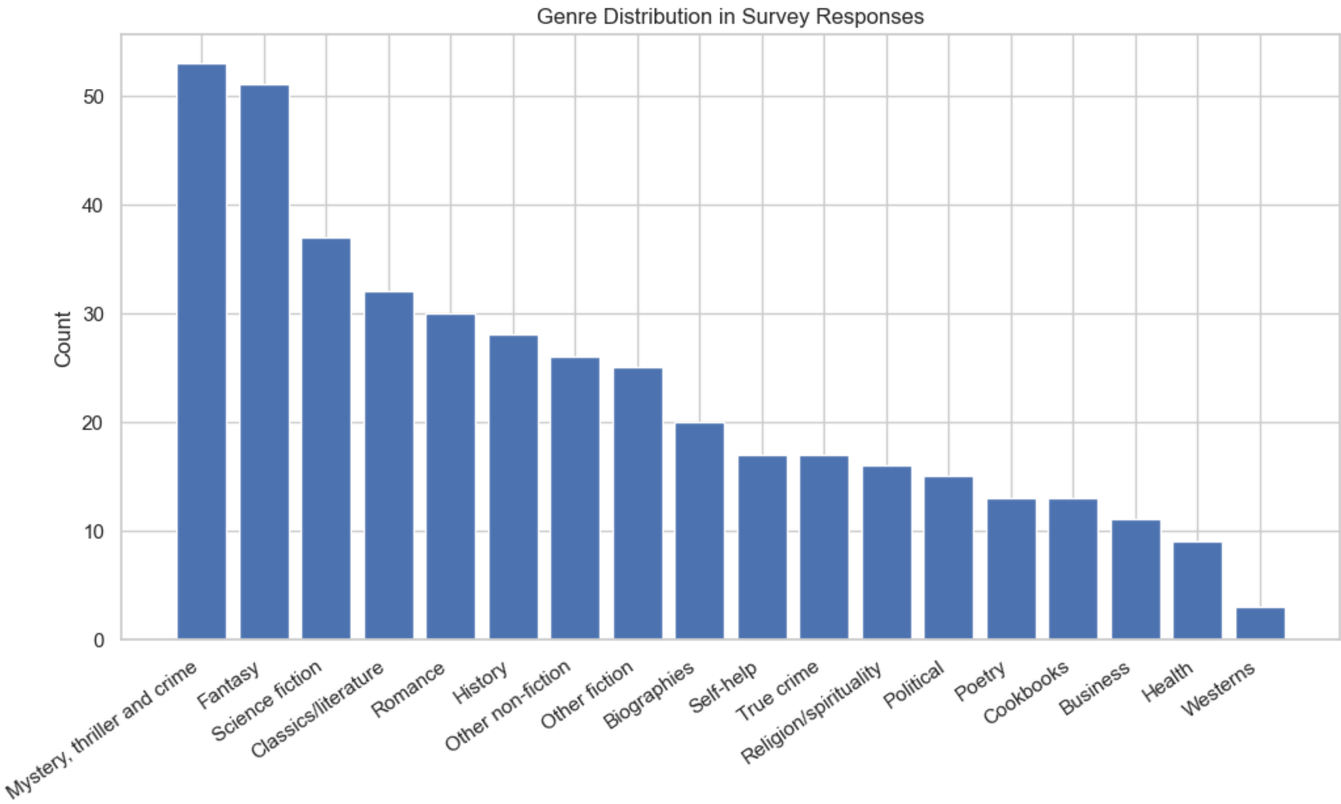


Figure 4.3: Genre distribution sorted by net occurrences.

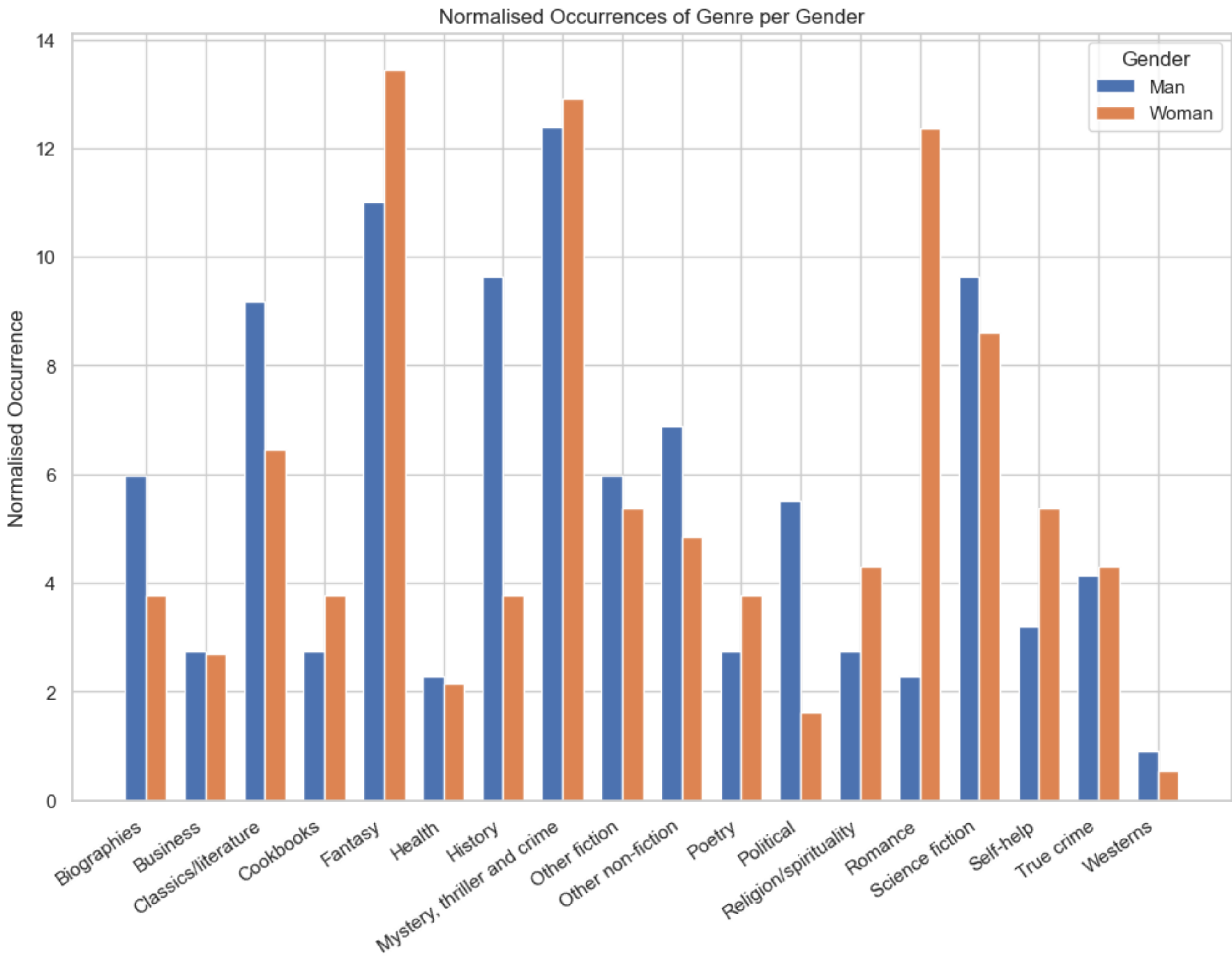
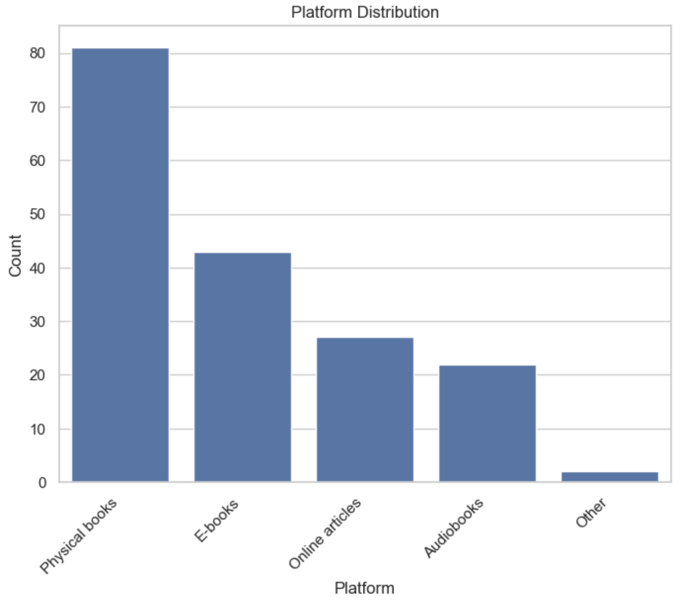
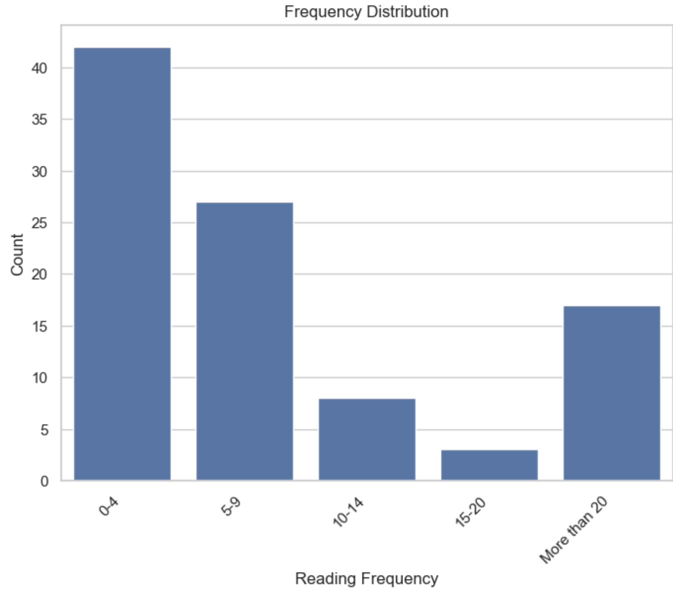


Figure 4.4: Normalised genre distribution by gender.



(a) Preferred reading platform distribution.



(b) Books consumed per year.

Figure 4.5: Preferred reading platforms and books consumed per year distribution for the experiment participants.

4.2 Response Model

To answer the first research question regarding the factors influencing the users' initial impression of the algorithms, the research findings underwent confirmatory factor analysis (CFA) and structural equation modelling (SEM). We used Lavaan [77] for R [71] to compute the CFA and SEM, treating all responses to the questions as ordinal variables. Each question was assigned to the factor it was intended to measure. Table 4.2 displays the factor loadings for each question for the initial CFA and the resulting simplified SEM.

In the complete CFA, there were a few questions with very low explanatory power. These include, for example: 'Which recommender more represents mainstream tastes instead of your own?' for understands me with $R^2 = 0.001$. Additionally, a high correlation was found among the factors Accuracy, Understands Me, and Satisfaction (correlation coefficients above 0.95). Thus, it cannot be assumed that these factors measure different aspects of the experiment.

For the simplified SEM, we removed the factors Satisfaction and Understands Me, retaining Accuracy as the factor. We initially experimented with keeping Satisfaction and Understands Me since they had higher explanatory factors, but Accuracy yielded better results in the end. Additionally, all four questions for Accuracy loaded strongly. Poorly-loading

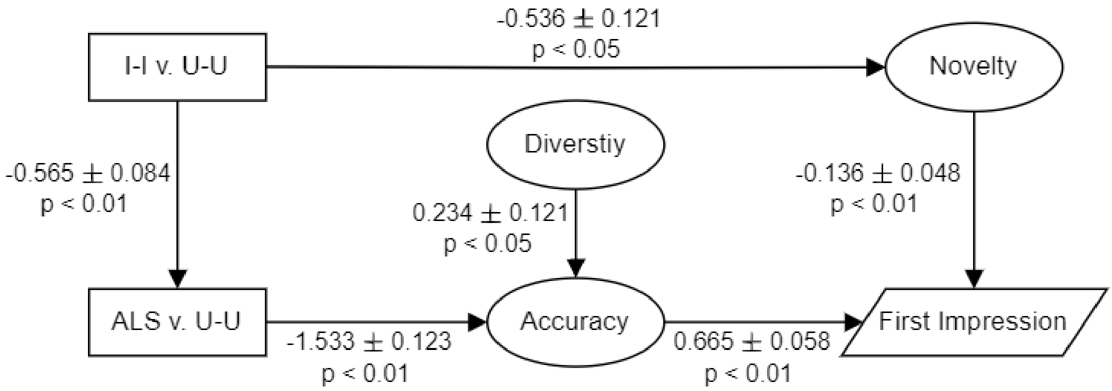


Figure 4.6: SEM with standard errors. The baseline condition is I-I vs. ALS

questions from Diversity ('Which list has more books that are similar to each other?' R^2 = 0,036) and Novelty ('Which list provides fewer new suggestions?' R^2 = 0,001) were removed.

We expanded the SEM by adding structural properties between factors, regressing them against the experimental conditions, and regressing the user's first impression against the experimental factors [26]. The final SEM was built by initially creating a large SEM that included many possible relationships, even those that theoretically make little sense. Then, we removed relationships that were not statistically significant. The result was a model that depicts all direct and indirect relationships. For example, the influence of Diversity on the first impression can be explained by its influence on Accuracy. The full lavaan output for the CFA and SEM can be viewed in Appendix B.

The results of the model mostly confirm the hypothesis that we expect similar results to the research by Ekstrand et al. Figure 4.6 shows the SEM with standard errors and the significance of the coefficients. We used Item-Item vs. ALS as the baseline algorithm and included the other conditions (Item-Item vs. User-User and ALS vs. User-User) as dummy variables. The Lavaan code used to create the SEM can be viewed in Listing 4.2 and the code for the CFA in Listing 4.1.

Listing 4.1: Lavaan model for the CFA

```
Acc =~ NA * AccAppealing + AccBest + AccBad + AccAtTop
Div =~ NA * DivSimilar + DivVaried + DivMoods + DivTastes
Und =~ NA * UndTaste + UndTrust + UndPersonalized + UndMainstream
Sat =~ NA * SatFind + SatRecommend + SatValuable + SatMobile + SatSat
Nov =~ NA * NovUnexpected + NovFamiliar + NovSurprising + NovUnthought +
    NovFewerNew
Acc ~~ 1*Acc
```

```
Div ~~ 1*Div
Nov ~~ 1*Nov
Sat ~~ 1*Sat
Und ~~ 1*Und
```

Listing 4.2: Lavaan model for the final SEM

```
Div =~ NA * DivVaried + DivMoods + DivTastes
Acc =~ NA * AccAppealing + AccBest + AccBad + AccAtTop
Nov =~ NA * NovUnexpected + NovFamiliar + NovSurprising + NovUnthought
Div ~~ 1*Div
Acc ~~ 1*Acc
Nov ~~ 1*Nov

CondALSUU ~ CondIIUU

Acc ~ Div + CondALSUU
Nov ~ CondIIUU

FirstImpression ~ Acc + Nov
```

Table 4.2: Response distributions: Navy entries have significant bias $p < 0.01$, sky blue no bias (uncorrected Wilcox test)

Factor / Question (W. l. = 'Which list', W. r. = 'Which recommender')	II v. ALS	II v. UU	ALS v. UU	Full CFA Coef.	Full CFA R^2	SEM Coef.
First Impression						
Accuracy					**0.74**	
W. l. has more books that you find appealing?				0.925	0.86	0.713
W. l. has more books that might be among the best books you read in the next year?				0.899	0.81	0.731
W. l. has more obviously bad book recommendations for you?				-0.720	0.52	-0.567
W. r. does a better job of putting better books at the top?				0.778	0.61	0.616
Diversity					**0.61**	
W. l. has more books that are similar to each other?				0.190	0.04	
W. l. has a more varied selection of books?				-0.703	0.49	0.657
W. l. has more books that match a wider variety of moods?				-0.814	0.66	0.863
W. l. would suit a broader set of tastes?				-0.619	0.38	0.609
Understands Me					**0.84**	
W. r. better understands your taste in books?				0.925	0.86	
W. r. would you trust to provide you with recommendations?				0.914	0.84	
W. r. seems more personalised to your book ratings?				0.909	0.83	
W. r. more represents mainstream tastes instead of your own?				0.036	0.00	
Satisfaction					**0.95**	
W. r. would better help you find books to read?				0.892	0.80	
W. r. would you be more likely to recommend to your friends				0.862	0.74	
W. l. of recommendations do you find more valuable?				0.880	0.77	
W. r. would you rather have as an app on your mobile phone?				0.885	0.78	
W. r. would better help to pick satisfactory books?				0.917	0.84	
Novelty					**0.04**	
W. l. has more books you do not expect?				0.490	0.24	0.671
W. l. has more books that are familiar to you?				-0.841	0.71	-0.600
W. l. has more pleasantly surprising books?				-0.528	0.28	-0.461
W. l. has more books you would not have thought to consider?				0.518	0.27	0.698
W. l. provides fewer new suggestions?				-0.029	0.00	

4.3 RQ1: Factors on First Impression

To answer research question 1 regarding the driving factors in the user's first impression of the algorithms, we examine the influence of the factors of Novelty, Accuracy, and Diversity on the user's first impression.

Directly influencing the decision are the factors of Novelty (negatively) and Accuracy (positively). Users seem to prefer algorithms with a perceived high accuracy. Additionally, presenting too many novel books is detrimental to the user's first impression. Furthermore, Diversity influences accuracy and thus indirectly affects the first impression positively. It is, therefore, crucial to consider the diversity of recommendations and not just focus on accuracy.

4.4 RQ2: Algorithm Performance

In Research Question 2, we aim to investigate the extent to which the subjective perception of individual factors varies between the algorithms. As discussed earlier, Table 4.1 depicts the first impression performance of the three algorithms: ALS clearly outperforms both algorithms, and Item-Item lags behind User-User in case a preference is identified.

Although the constructed Structural Equation Model (SEM) incorporates experimental conditions with dummy variables, the interpretation of the conditions is challenging due to the comparative structure of the study. To scrutinize the differences among the algorithms more precisely, we examine the response distribution of each question in the study for the algorithm pairings, illustrated in Table 4.2. A significant bias ($p < 0.01$) was observed in distributions in navy blue using an uncorrected Wilcoxon test. No bias was detected in distributions in sky blue.

Regardless of which algorithms were evaluated, it can be observed that there was no significant difference in the perceived diversity under any condition. This is also confirmed by the objective metric of diversity. This will be explored in more detail in section 3.8. After this general observation, let's take a closer look at the individual pairings:

Item-Item vs. ALS

Overall, not many significant differences were found between the two algorithms. However, ALS seems to suggest more appealing books. Additionally, ALS was perceived to better capture the reader's taste and provide more familiar books. In summary, it can be concluded that ALS generates recommendations with higher perceived accuracy and better perceived personalisation. On the other hand, Item-Item generates more novel recommendations.

These findings are supported by the objective metrics of both algorithms.

Item-Item vs. User-User

In this pairing, more significant differences were identified. Users perceived that Item-Item recommended obviously bad books significantly more frequently. This overall resulted in a lower perceived accuracy of Item-Item recommendations. Additionally, the recommendations from User-User were perceived as more personalised. Furthermore, the overall user satisfaction with User-User lists was higher, as measured by two significantly different questions. The differences in novelty are also significant: Item-Item, again, provides fewer familiar books, while User-User recommends more pleasantly surprising books. The objective metric of accuracy for Item-Item was slightly lower than that of User-User, contradicting the investigation of subjective user perception. The same applies to the novelty factor, where the extent of the difference is greater.

ALS vs. User-User

In this experimental condition, by far the most significant differences were observed. In a total of 13 out of 22 questions, a significant difference was observed, with ALS consistently being the preferred algorithm. ALS is significantly rated better in terms of accuracy across all four questions. The same observation holds for the perceived personalisation of recommendations and the overall user satisfaction with the generated recommendation list. The only factor, besides diversity, that was not as decisively in favour of ALS was novelty. Here, only one significant difference was measured: ALS provides more familiar books in its recommendations. No significant difference was found in the other questions. The observations regarding accuracy and novelty can be supported by the respective objective metrics.

4.5 RQ3: Bias in different User Groups

To answer the question of how various characteristics of user groups influence the examined factors of the experiment, we expanded the SEM from section 4.2 to include collected demographic data. This involved examining the influence of personal data such as gender and age, as well as domain knowledge. The latter was determined using an 'expert score'.

The expert score is composed of both the stated annual reading frequency and the preferred genres for each user. The dataset [41], which contains textual descriptions and

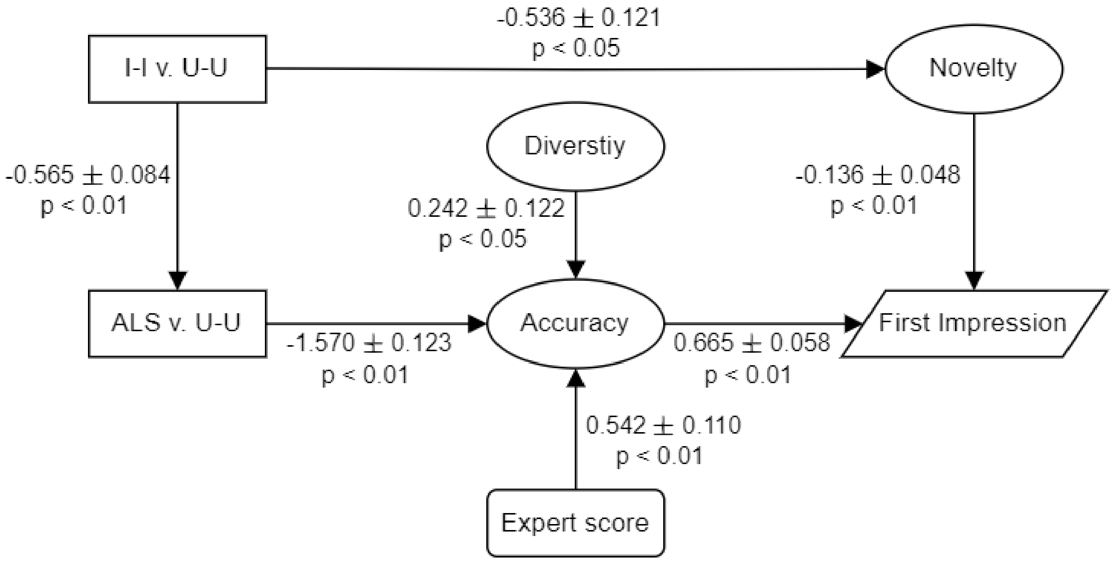


Figure 4.7: SEM expanded with demographic data. The baseline condition is I-I vs. ALS

genres for 10,000 books, served as the basis for this calculation. Using Natural Language Processing (NLP) and cosine similarity, a similarity score was calculated for each genre, indicating how similar each genre is to science fiction[1]. The expert score for a user is calculated by summing all similarity scores (as each user could specify multiple preferred genres), normalising by the total number of preferred genres, and finally multiplying by a fixed factor. This factor is determined by the annual reading frequency, where '0-4' corresponds to a factor of 1, '5-9' to a factor of 2, and finally 'More than 20' to a factor of 5. The remaining frequencies are determined with the same pattern. Different age groups were added to the SEM by interpreting the values as ordinal variables. Gender was included by considering 'Woman' as the baseline and including 'Man' as a dummy variable. The code for the expanded model can be viewed in Listing 4.3.

No significant influences of gender and age on any factor were observed. This contradicts the assumption in section 3.7 that age and gender have a significant impact on the perceived accuracy of the algorithms. However, a significant ($p < 0.01$), strong, positive (0.542) influence of the expert score on accuracy was identified. Thus, this experiment at least partially supports Knijnenburg et al. statement that age, gender, and domain knowledge can have an impact on various factors [45]. The modified SEM is depicted in Figure 4.7. The full lavaan output for the expanded SEM can be viewed in Appendix B.

[1]The similarity to science fiction was calculated, since the algorithms provided recommendations for a science fiction fan.

Listing 4.3: Lavaan model for the expandend SEM with the expert score

```
Div =~ NA * DivVaried + DivMoods + DivTastes
Acc =~ NA * AccAppealing + AccBest + AccBad + AccAtTop
Nov =~ NA * NovUnexpected + NovFamiliar + NovSurprising + NovUnthought
Div ~~ 1*Div
Acc ~~ 1*Acc
Nov ~~ 1*Nov

CondALSUU ~ CondIIUU

Acc ~ Div + CondALSUU + expert_score
Nov ~ CondIIUU

FirstImpression ~ Acc + Nov
```

4.6 RQ4: Objective Measures

To answer research question 4, we examine the relationship between the objective metrics and the perceived preferences of users. In Figure 4.8, all distributions of the calculated objective metrics are shown.

The assessment that ALS provides the lists with the highest accuracy, i.e., making the least errors in predicting ratings, is supported by the RMSE score. However, the extent of the objective metric is not as drastic as suggested by user survey responses. The objective accuracy of User-User and Item-Item is close to each other, reflecting the subjective perception. There was only one significant difference in one of the questions addressing accuracy between the two algorithms.

As mentioned in section 3.8, values for Intra-List Similarity must be interpreted with caution, as similarity was not calculated for all books in the respective lists. Values were missing for 3 (User-User, ALS) or 5 (Item-Item) books in the lists. There is a tendency that the intra-list similarity score does not differ much between User-User and ALS, indicating that both algorithms deliver similar diverse lists, which is supported by the user responses. When looking at the complete subjective perception of users regarding diversity, one can speculate that with more tag data, the objective score of Item-Item will become similar to that of the others, as users did not perceive significant differences between all 3 algorithms. This assumption needs to be verified in the future with a more robust tag-genome similarity analysis.

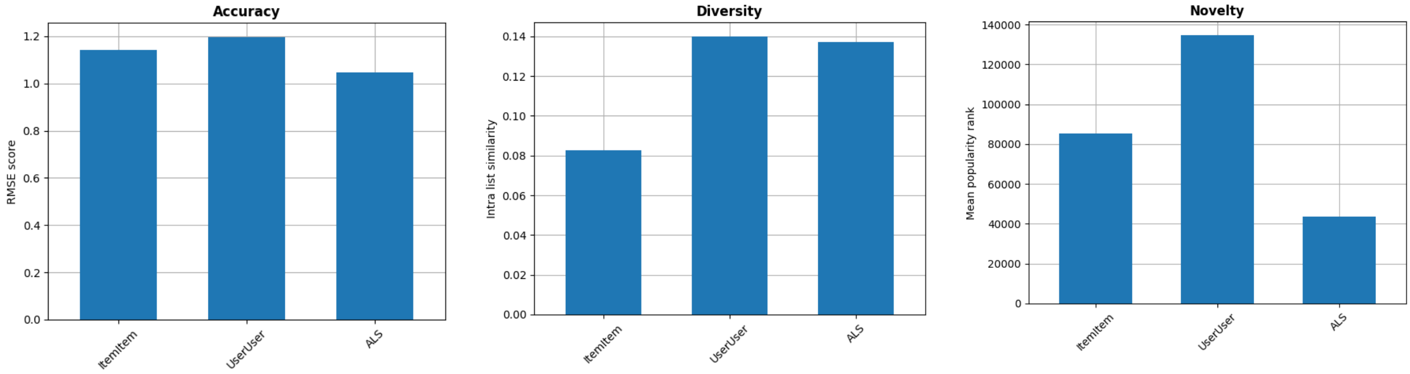


Figure 4.8: Objective measures of the observed algorithms.

Novelty was the only factor where significant differences were found between the objective metrics of the algorithms. The distribution of the mean-popularity-rank shows that User-User generates lists with less popular (likely more novel/unusual) items than ALS and Item-Item. ALS generates lists with more well-known books. This observation is only partially supported by users' subjective perceptions. The assumption regarding ALS is confirmed in both conditions where ALS was part of the observation through the question 'Which list has more books that are familiar to you.' In both cases, this question is the only one with significant relevance, and in both cases, the decision is skewed towards ALS. The objective observation that User-User provides more obscure lists is not confirmed by user responses. In the condition of User-User vs. Item-Item, the books of the Item-Item list were perceived as more novel.

In summary, the trends of the objective metrics align with those of the subjective perception of users in almost all cases (excluding diversity of Item-Item due to data limitations). The only significant difference was the 'swapped' perception of Novelty of User-User and Item-Item considering the objective metrics.

4.7 RQ5: Logging Patterns

To answer research question 5 and gain insights into user behaviour with the experiment interface, we are examining the collected log data in more detail. The method preferred by users to obtain additional information about the displayed books was the integration of the 'Description Buttons'. These were clicked a total of 909 times, meaning that each user, on average, viewed approximately 9 descriptions. The description for 'The Doors of Time' was the most accessed (57 clicks), while the one for 'Harry Potter and the Prisoner of Azkaban' was clicked the least (15 times). The latter is easily explained since almost everyone is familiar with the book series, and they roughly know what it is about.

The information about clicks enhances the quality of the collected study data, indicating that users engaged extensively with the displayed content. Furthermore, it demonstrates that information not immediately present but accessible with a click within the same interface is well-received by users.

The interaction with elements that display information on another page, as observed with the Clicks on the ‘Cover Images‘, is less frequent. In total, only 143 clicks, about 15% of the description button clicks, were recorded. This could be attributed to the fact that this functionality was more ‘hidden‘ compared to an obvious button. Additionally, as mentioned earlier, the information opened in a new tab rather than being displayed directly on the website, which might have influenced user engagement. Table 4.3 shows the distribution of clicks on the cover image and description button for each book present in the study.

The Back to Survey button explained in section 3.2, which allowed users to participate in the study again after successful completion, was well-received. In total, there were 43 clicks recorded. This means that approximately 44% of all participants expressed interest in seeing other configurations of recommendation lists. Nine out of these 43 users subsequently completed the entire questionnaire again successfully for a different combination of lists. Thus, every 8th participant participated in the study multiple times.

Therefore, it is at least worth considering incorporating such a mechanism when multiple participations make sense (provided that the content of the study changes). A user who has already participated is familiar with the study and knows what to expect, potentially leading to more qualitative data.

Book title	Button clicks	Cover clicks
Losing Connor	55	10
Free to Be... You and Me	39	6
A Shot at the Big Time	29	6
We, the Drowned	35	4
The Grapes of Wrath	48	4
Beartown	20	3
Stoner	26	5
Self-Publishing Steps To Successful Sales	12	2
Strangers in Paradise	14	5
Chicago Manual of Style	23	3
So Long, and Thanks for All the Fish	53	15
The Doors of Time	57	7
The Hitchhiker's Guide to the Galaxy	51	7
The Duchess War	35	5
At Least There's The Football	40	4
The Grapes of Wrath	48	4
Act Like It	35	4
The Wind-Up Bird Chronicle	27	3
Attack of the Deranged Mutant Killer Monster Snow Goons	21	4
The Winter King	22	6
The Martian	32	4
Thirteen	48	5
Ready Player One	37	4
An Amish Home	33	4
The Wake	31	4
The Hitchhiker's Guide to the Galaxy	51	7
Jesus the Christ	20	6
Harry Potter and the Prisoner of Azkaban	10	2
The Black Eagle Inn	29	4
Curse of Strahd	27	7
Total	909	143

Table 4.3: Description button and cover image click per each book divided by algorithm (Item-Item = red, User-User = yellow, ALS = blue).

Chapter 5

Discussion

This study serves as a continuation of Ekstrand et al.'s study on movie recommenders [26] to extend the key insights from the widely studied film domain to the book domain and investigate whether similar findings can be observed there. We were able to expand our understanding of user preferences regarding book recommendation systems through the use of structural equation modelling (SEM). The fundamental aspects influencing the initial impression of book recommendation systems were novelty (negative) and accuracy (positive). Additionally, domain knowledge has positively influenced the perception of accuracy. By analyzing the response distributions of each question in the study's questionnaire, we were able to determine which algorithm achieved the highest user satisfaction and rating. In our study, this was the ALS algorithm, which was consistently better evaluated in both cases where it was part of the configuration. The magnitude of this effect was greater in the User-User comparison than in the Item-Item comparison. Therefore, ALS can be considered the winner in this experiment, as the generated results best-met user expectations, and the recommendations were most relevant to their interests. Additionally, we analyzed the logging data collected during the experiment to identify trends and similarities in user behaviour with the experimental interfaces. We discovered that components displaying additional information, in our case, more book details, on the same page were better received than those outsourcing the information to a new browser window. This information should be considered in future similar experiments during the design phase to enhance the effectiveness of the interface. Furthermore, we recognized the potential benefits of incorporating a feature allowing users to participate in the study multiple times, if experiment design permits multiple participation. This function was well-received by the participants.

In the following, we will discuss our key findings in more detail and will compare them, if applicable, with similar previous studies like [26] and [45].

5.1 Effects of Novelty

The significant negative effect of the novelty of recommended books and the accompanying loss of user trust in the system, that the algorithm can not provide satisfactory results, was particularly noticeable. This effect was especially pronounced when assessing the algorithms' first impressions.

This implies that designers of recommendation systems need to be mindful not to recommend too many novel items initially. Too many unknown items can be detrimental to user perception. It would be interesting to investigate in the domain of book recommendations whether the negative effect of novelty diminishes for experienced users of a system who have been using it for a longer time.

These findings complement those of Ekstrand et al. They also observed a strong negative influence of novelty on the user's first impression of recommendation lists. Additionally, they found a significant negative impact of novelty on user satisfaction and perceived accuracy, which we could not confirm in our study. This implies that the negative effect of novelty is not necessarily domain-specific and may have general significance for recommendation systems. This supports the assertion that, in the design of recommendation algorithms, one should be cautious about novelty to prevent these negative effects. The results from [60] also extend the negative impact of novelty in recommendations on perceived personalisation, indicating how well the algorithm recognizes personal preferences and suggests items accordingly. Stating, that more novelty comes at the cost of lower perceived preference fit. Furthermore, these findings confirm that it is crucial for user satisfaction to establish a certain level of trust in the recommendation system from the beginning, and the generation of too many unfamiliar items can be detrimental to that goal [26, 62].

5.2 Effects of Diversity

Regarding perceived diversity, it was found that it had a significant positive impact on perceived accuracy, which, in turn, indirectly leads to a better first impression. This means, that even if diversity does not have a direct impact on the first impression of the algorithm, it still should be carefully considered when designing a recommender algorithm due to the positive impact on the accuracy. The observations mentioned section 2.3 that generating more diverse lists may compromise accuracy [99, 105, 107] and many diversification techniques result in traditional objective metrics for accuracy yielding poorer results are very interesting in light of our results. The positive impact of diversity on accuracy in our study suggests that such a negative relationship between the two factors does not exist in user perception and that this effect is more reflected in traditional objective metrics. This effect is supported by Ektrand et al.'s results. They also found a significant influence of diversity on user satisfaction[1]. Additionally, a mild but positive influence of novelty on diversity was measured there, but it did not occur in our study. The findings of the relationship between accuracy and diversity further support the hypothesis, that the negative impact measured in previous studies does not hold in the analysis of user perception.

Finally, it can be observed that the objective measurement of diversity through tag genome similarity, proposed by Ektrand et al., seems to be applicable to the book domain as well. Our study, using a limited tag genome dataset, shows that initial trends provide credible results. However, to make a more accurate statement, the calculation of such a diversity metric needs to be carried out using a larger dataset where not as many entries of relevant books are missing.

5.3 Effects of Accuracy

It was found that the factors Accuracy, Understands Me, and Satisfaction all strongly correlate with each other, suggesting that they may not measure different aspects of user perception. One can therefore, with some reservations, equate these factors. This could mean either that these factors simply do not describe different aspects in the analysis of user preferences, or that the selected questions in the questionnaire used were not optimal for measuring the differences in these factors. To answer this question, further research needs to build upon and investigate whether using different questions, these factors address different aspects of user perception. In Ekstrand et al.'s research, the same relationship between the three factors can be observed. Here too, Accuracy, Understands Me, and Satisfaction are strongly correlated. This suggests that this observation is not domain-specific and may be generally applicable.

Furthermore, it was found that perceived accuracy has a strong positive impact on the first impression of a recommendation list. Users seem to prefer recommendation lists that have been accurately generated. This observation is also supported by Ekstrand et al.'s results, where the same effect of Accuracy was observed. In their study, Accuracy was a significantly positive factor for the first impression of the lists. These results are further supported by the research findings of [44], which found that in traditional recommendation algorithms, such as collaborative-filtering algorithms used in our study, perceived accuracy positively influences user satisfaction.

[1] User satisfaction and perceived accuracy can be equated in this context, as both studies have shown a strong correlation between these two factors.

5.4 Algorithm Performance

Regarding the performance of individual algorithms, it was observed that no significant difference could be found in terms of the diversity factor. Neither in subjective perception nor, considering the lack of data (more details in Figure 4.8), in the objective metrics was a significant difference observed. Furthermore, ALS clearly outperformed User-User in the comparison of the subjective user data, with a positive preference for ALS found in almost all questions. The comparison of ALS with Item-Item is not as clear-cut. The two algorithms performed roughly equally, with ALS being preferred in only one question for each of three out of five factors. This suggests that although the two algorithms were perceived roughly similarly, ALS performs slightly better than Item-Item in most factors. Based on these two results, that ALS clearly wins over User-User and is only slightly preferred compared to Item-Item, the observation that User-User wins the comparison with Item-Item is surprising. It should be assumed that User-User loses the comparison with Item-Item, but this is not the case. User-User is preferred in all four factors (excluding diversity) in one or two questions. The question then arises as to why User-User loses so clearly to ALS but wins against Item-Item in turn. A more extensive study is necessary to investigate whether this observation also appears with a larger sample size.

As for the distribution of the first impression of recommendation lists, ALS clearly wins against both algorithms. User-User wins over Item-Item, albeit less decisively, if a preference is observed. In the latter case, it should be mentioned again that there were also some assessments indicating that the two algorithms generated lists of roughly equal attractiveness. Unfortunately, the comparative design of the experiment does not allow us to determine how well or poorly these lists performed.

In terms of algorithm performance in various objective metrics, no significant differences were found in accuracy. All algorithms had similar RMSE values. Nevertheless, ALS was significantly more frequently preferred in the first impression, indicating, to a degree, that offline metrics do not capture what ultimately influences the user experience with a recommendation system.

The only significant difference in objective metrics was observed in novelty. Here, ALS provides by far the most familiar recommendation lists, and User-User provides the least familiar ones. Item-Item falls roughly in the middle of the two algorithms. All three tendencies in offline metrics can be confirmed by the results of Ekstrand et al. The distributions of both studies have a similar shape. An exception is the difference in the obscurity metric. Here, the general difference, but especially that between Item-Item and ALS (respectively SVD), is not as extreme as in our study.

5.5 Bias in User Groups

The investigation of potential bias among different user groups revealed that users' domain knowledge had a significantly positive impact on the perceived accuracy of recommendation lists. Users with expertise in the science fiction book domain found it important for algorithms to generate accurate recommendations. No other effects of the calculated expert score on the other factors were observed. It would be interesting to explore other ways of creating a score that measures users' domain knowledge and potentially identify differences in its impact on user perception.

For other demographic data such as gender and age group, no influences on any of the 3 (respectively 5) factors were found. This should be further investigated in future studies, as research by Knijnenburg et al. [45] has shown that age, gender, and domain knowledge can influence factors in user perception. Only the influence of domain knowledge was demonstrated in our study. One possible explanation for the absence of the age effect could be related to the unbalanced age distribution of the study participants. The participants were predominantly young adults, with other age groups being underrepresented. A study on the influences of participants' age on the user perception of the algorithms should ideally be conducted again with a balanced age distribution.

Another interesting investigation, assuming that participants are demographically roughly evenly distributed, would be whether different demographic groups significantly prefer different algorithms over ones chosen by another group varying in demographical attributes.

5.6 Limitations and Generalisability

An obvious limitation of the experiment lies in the nature of the research scenario and the resulting generation of recommendations. Unlike the study by Ekstrand et al., recommendations were not personalized for each individual participant in our study but were generated for the user in the research scenario. This limitation imposes a significant demand on participants, who had to empathise with the role of another person. While efforts were made to implement this as effectively as possible by recruiting participants with thematic relevance, an optimal condition for such a study cannot be fully achieved. It would be interesting to conduct such a study on an existing book recommendation platform like Goodreads or similar, both to gain access to historical user data for the purpose of generating personalised recommendations and to have a broad pool of potential participants who possess extensive knowledge in the domain of books.

Another limitation arises from the comparative nature of the experiment. While nuances in the differences between algorithms can be identified, it cannot be measured, for example, whether the algorithms were equally good or equally poor. The results should, therefore, be contextualised with studies that have adopted a non-comparative design to gain a deeper understanding of the algorithms.

Furthermore, through this study and the previous one by Ekstrand et al., two domains have been investigated, yielding similar results. This is a step towards obtaining a general assessment of user perception of recommendation systems, but more studies from different domains are needed to confirm these findings.

Additionally, the additional questions regarding the influence of demographic differences among user groups on user perception need further exploration.

Ekstrand et al.'s study provides a good starting point for similar analyses in different domains, as our study has demonstrated.

Chapter 6

Conclusion

This study served as a follow-up experiment to Ekstrand et al. [26], comparing the output of three widely used collaborative filtering algorithms and analyzing subjective differences perceived by users. We transferred the investigations of the previous study to the book domain, achieving comparable results. This expanded the understanding of the influences of diversity, novelty, and accuracy on user reception of recommendation systems to another domain, representing the next step toward drawing general conclusions in the field of user perception of recommendation systems. Additionally, our study extended the analysis to include the influences of demographic data, shedding light on the roles of gender, age, and user expertise in relation to the user perception of recommendation lists.

There are many interesting possible extensions of this research. Firstly, similar research should be conducted in other domains to examine the extent to which the results confirmed by now two studies are applicable to different domains. Secondly, the studied algorithms could be expanded to include a representative of neural network collaborative filtering algorithms, as they exhibit the promising characteristic of generating more diverse recommendations [5]. Another avenue for further research is the incorporation of additional factors that describe different aspects of user perception. Notably, the factor of serendipity, which in other studies, unlike novelty, had a significantly positive impact on user satisfaction [15, 44]. Constructing suitable questions that capture this factor but are not too similar to novelty questions can be a challenging but insightful task. A final potential extension could be the integration of such a study into an existing book recommendation platform, which would open up several avenues. On the one hand, personalised recommendations could be created for participants, and on the other hand, there would be a larger pool of both demographic data and response data, allowing for more robust conclusions.

Understanding how users perceive different recommendation algorithms is necessary to develop and design new algorithms that are better tailored to user needs and not rigidly trained using objective metrics. This study demonstrates that the results of Ekstrand et al. were not domain-specific and suggests that these findings can be helpful in a broad range of applications.

Bibliography

[1] Himan Abdollahpouri et al. “Toward the Next Generation of News Recommender Systems”. In: *Companion Proceedings of the Web Conference 2021*. WWW ’21. Ljubljana, Slovenia: Association for Computing Machinery, 2021, pp. 402–406. ISBN: 9781450383134. DOI: 10.1145/3442442.3452327. URL: https://doi.org/10.1145/3442442.3452327.

[2] G. Adomavicius and A. Tuzhilin. “Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions”. In: *IEEE Transactions on Knowledge and Data Engineering* 17.6 (2005), pp. 734–749. DOI: 10.1109/TKDE.2005.99.

[3] Muhammad Ali et al. “Discrimination through Optimization: How Facebook’s Ad Delivery Can Lead to Biased Outcomes”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359301. URL: https://doi.org/10.1145/3359301.

[4] *Amazon*. [Online, accessed 03-November-2023]. URL: https://www.amazon.de/.

[5] Vito Walter Anelli et al. “Reenvisioning the Comparison between Neural Collaborative Filtering and Matrix Factorization”. In: *Proceedings of the 15th ACM Conference on Recommender Systems*. RecSys ’21. Amsterdam, Netherlands: Association for Computing Machinery, 2021, pp. 521–529. ISBN: 9781450384582. DOI: 10.1145/3460231.3475944. URL: https://doi.org/10.1145/3460231.3475944.

[6] Poonam B.Thorat, R. Goudar, and Sunita Barve. “Survey on Collaborative Filtering, Content-based Filtering and Hybrid Recommendation System”. In: *International Journal of Computer Applications* 110 (Jan. 2015), pp. 31–36. DOI: 10.5120/19308-0760.

[7] Javad Basiri et al. “Alleviating the cold-start problem of recommender systems using a new hybrid approach”. In: *2010 5th International Symposium on Telecommunications*. 2010, pp. 962–967. DOI: 10.1109/ISTEL.2010.5734161.

[8] D. Bawden and L. Robinson. “Information Overload: An Overview”. In: *Oxford Encyclopedia of Political Decision Making*. This is a draft of a chapter that has been published by Oxford University Press in the book Oxford Encyclopedia of Political Decision Making. Oxford: Oxford University Press, June 2020. DOI: 10.1093/acrefore/9780190228637.013.1360. URL: https://openaccess.city.ac.uk/id/eprint/23544/.

[9] Daniel Billsus and Michael J. Pazzani. “Learning Collaborative Information Filters”. In: *Proceedings of the Fifteenth International Conference on Machine Learning*. ICML ’98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 46–54. ISBN: 1558605568.

[10] Dimitrios Bountouridis et al. “SIREN: A Simulation Framework for Understanding the Effects of Recommender Systems in Online News Environments”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 150–159. ISBN: 9781450361255. DOI: 10.1145/3287560.3287583. URL: https://doi.org/10.1145/3287560.3287583.

[11] Brian. *Readers Also Enjoyed...* [Online, accessed 10-November-2023]. May 2011. URL: https://www.goodreads.com/blog/show/282-readers-also-enjoyed.

[12] Robin Burke. “Hybrid Recommender Systems: Survey and Experiments”. In: *User Modeling and User-Adapted Interaction* 12 (Nov. 2002). DOI: 10.1023/A:1021240730564.

[13] Robin Burke, Alexander Felfernig, and Mehmet H. Göker. “Recommender Systems: An Overview”. In: *AI Magazine* 32.3 (June 2011), pp. 13–18. DOI: 10.1609/aimag.v32i3.2361. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2361.

[14] Erion Çano and Maurizio Morisio. “Hybrid recommender systems: A systematic literature review”. In: *Intelligent Data Analysis* 21 (2017). 6, pp. 1487–1524. ISSN: 1571-4128. DOI: 10.3233/IDA-163209. URL: https://doi.org/10.3233/IDA-163209.

[15] Li Chen et al. “How Serendipity Improves User Satisfaction with Recommendations? A Large-Scale User Evaluation”. In: *The World Wide Web Conference*. WWW ’19. San Francisco, CA, USA: Association for Computing Machinery, 2019, pp. 240–250. ISBN: 9781450366748. DOI: 10.1145/3308558.3313469. URL: https://doi.org/10.1145/3308558.3313469.

[16] Heng-Tze Cheng et al. "Wide & Deep Learning for Recommender Systems". In: *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*. DLRS 2016. Boston, MA, USA: Association for Computing Machinery, 2016, pp. 7–10. ISBN: 9781450347952. DOI: 10.1145/2988450.2988454. URL: https://doi.org/10.1145/2988450.2988454.

[17] Alexander Chernev, Ulf Böckenholt, and Joseph Goodman. "Choice overload: A conceptual review and meta-analysis". In: *Journal of Consumer Psychology* 25.2 (2015), pp. 333–358. ISSN: 1057-7408. DOI: https://doi.org/10.1016/j.jcps.2014.08.002. URL: https://www.sciencedirect.com/science/article/pii/S1057740814000916.

[18] Muhammad Waseem Chughtai et al. "E-Learning Recommender Systems Based on Goal-Based Hybrid Filtering". In: *International Journal of Distributed Sensor Networks* 2014 (July 2014). DOI: 10.1155/2014/912130.

[19] Kyusik Chung. *Announcing Goodreads Personalized Recommendations*. [Online, accessed 10-November-2023]. Sept. 2011. URL: https://www.goodreads.com/blog/show/303-announcing-goodreads-personalized-recommendations.

[20] Mukund Deshpande and George Karypis. "Item-Based Top-N Recommendation Algorithms". In: *ACM Trans. Inf. Syst.* 22.1 (Jan. 2004), pp. 143–177. ISSN: 1046-8188. DOI: 10.1145/963770.963776. URL: https://doi.org/10.1145/963770.963776.

[21] PV Devika et al. "Book Recommendation System". In: *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. 2021, pp. 1–5. DOI: 10.1109/ICCCNT51525.2021.9579647.

[22] Benjamin Edelman, Michael Luca, and Dan Svirsky. "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment". In: *American Economic Journal: Applied Economics* 9.2 (Apr. 2017), pp. 1–22. DOI: 10.1257/app.20160213. URL: https://www.aeaweb.org/articles?id=10.1257/app.20160213.

[23] Michael D. Ekstrand. "The LKPY Package for Recommender Systems Experiments: Next-Generation Tools and Lessons Learned from the LensKit Project". In: *CoRR* abs/1809.03125 (2018). arXiv: 1809.03125. URL: http://arxiv.org/abs/1809.03125.

[24] Michael D. Ekstrand et al. "All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research.

PMLR, Feb. 2018, pp. 172–186. URL: https://proceedings.mlr.press/v81/ekstrand18b.html.

[25] Michael D. Ekstrand et al. “Letting Users Choose Recommender Algorithms: An Experimental Study”. In: *Proceedings of the 9th ACM Conference on Recommender Systems*. Vienna, Austria: Association for Computing Machinery, 2015, pp. 11–18. ISBN: 9781450336925. DOI: 10.1145/2792838.2800195. URL: https://doi.org/10.1145/2792838.2800195.

[26] Michael D. Ekstrand et al. “User Perception of Differences in Recommender Algorithms”. In: *Proceedings of the 8th ACM Conference on Recommender Systems*. RecSys ’14. Foster City, Silicon Valley, California, USA: Association for Computing Machinery, 2014, pp. 161–168. ISBN: 9781450326681. DOI: 10.1145/2645710.2645737. URL: https://doi.org/10.1145/2645710.2645737.

[27] Motahhare Eslami et al. “Be Careful; Things Can Be Worse than They Appear: Understanding Biased Algorithms and Users’ Behavior Around Them in Rating Platforms”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 11.1 (May 2017), pp. 62–71. DOI: 10.1609/icwsm.v11i1.14898. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14898.

[28] Simon Funk. *Netflix Update*. [Online, accessed 09-November-2023]. URL: https://sifter.org/simon/journal/20061211.html.

[29] Mustansar ali Ghazanfar and A. Prugel-Bennett. “An Improved Switching Hybrid Recommender System Using Naive Bayes Classifier and Collaborative Filtering”. In: *Lecture Notes in Engineering and Computer Science* 2180 (Mar. 2010).

[30] Carlos A. Gomez-Uribe and Neil Hunt. “The Netflix Recommender System: Algorithms, Business Value, and Innovation”. In: *ACM Trans. Manage. Inf. Syst.* 6.4 (Dec. 2016). ISSN: 2158-656X. DOI: 10.1145/2843948. URL: https://doi.org/10.1145/2843948.

[31] *Goodreads | Meet your next favorite book*. [Online, accessed 26-October-2023]. URL: https://www.goodreads.com/.

[32] Larry Hardesty. *The history of Amazon’s recommendation algorithm*. [Online, accessed 10-November-2023]. Nov. 2019. URL: https://www.amazon.science/the-history-of-amazons-recommendation-algorithm.

[33] Xiangnan He et al. “Neural Collaborative Filtering”. In: *CoRR* abs/1708.05031 (2017). arXiv: 1708.05031. URL: http://arxiv.org/abs/1708.05031.

[34] Jon Herlocker, Joseph Konstan, and John Riedl. “An Empirical Analysis of Design Choices in Neighborhood-Based Collaborative Filtering Algorithms”. In: *Information Retrieval* 5 (Jan. 2002), pp. 287–310. DOI: 10.1023/A:1020443909834.

[35] Christopher K. Hsee and Jiao Zhang. “General Evaluability Theory”. In: *Perspectives on Psychological Science* 5.4 (2010). PMID: 26162182, pp. 343–355. DOI: 10.1177/1745691610374586. eprint: https://doi.org/10.1177/1745691610374586. URL: https://doi.org/10.1177/1745691610374586.

[36] Liang Hu et al. “Personalized Recommendation via Cross-Domain Triadic Factorization”. In: *Proceedings of the 22nd International Conference on World Wide Web*. WWW ’13. Rio de Janeiro, Brazil: Association for Computing Machinery, 2013, pp. 595–606. ISBN: 9781450320351. DOI: 10.1145/2488388.2488441. URL: https://doi.org/10.1145/2488388.2488441.

[37] Yifan Hu, Yehuda Koren, and Chris Volinsky. “Collaborative Filtering for Implicit Feedback Datasets”. In: *2008 Eighth IEEE International Conference on Data Mining*. 2008, pp. 263–272. DOI: 10.1109/ICDM.2008.22.

[38] Basileal Imana, Aleksandra Korolova, and John S. Heidemann. “Auditing for Discrimination in Algorithms Delivering Job Ads”. In: *CoRR* abs/2104.04502 (2021). arXiv: 2104.04502. URL: https://arxiv.org/abs/2104.04502.

[39] Dietmar Jannach. “Evaluating conversational recommender systems”. In: *Artificial Intelligence Review* 56.3 (June 2022), pp. 2365–2400. DOI: 10.1007/s10462-022-10229-x. URL: https://doi.org/10.1007%2Fs10462-022-10229-x.

[40] Zhiyang Jia et al. “User-Based Collaborative Filtering for Tourist Attraction Recommendations”. In: *2015 IEEE International Conference on Computational Intelligence & Communication Technology*. 2015, pp. 22–25. DOI: 10.1109/CICT.2015.20.

[41] Ishika Johari. *Best Books (10k) Multi-Genre Data*. Last updated in April 2023 [Retrieved, 28-November-2023]. URL: https://www.kaggle.com/datasets/ishikajohari/best-books-10k-multi-genre-data.

[42] Diane Kelly. “Methods for Evaluating Interactive Information Retrieval Systems with Users”. In: *Foundations and Trends® in Information Retrieval* 3.1–2 (2009), pp. 1–224. ISSN: 1554-0669. DOI: 10.1561/1500000012. URL: http://dx.doi.org/10.1561/1500000012.

[43] Jaekyeong Kim, Ilyoung Choi, and Qinglong Li. "Customer Satisfaction of Recommender System: Examining Accuracy and Diversity in Several Types of Recommendation Approaches". In: *Sustainability* 13.11 (2021). ISSN: 2071-1050. DOI: 10.3390/su13116165. URL: https://www.mdpi.com/2071-1050/13/11/6165.

[44] Jaekyeong Kim, Ilyoung Choi, and Qinglong Li. "Customer Satisfaction of Recommender System: Examining Accuracy and Diversity in Several Types of Recommendation Approaches". In: *Sustainability* 13.11 (2021). ISSN: 2071-1050. DOI: 10.3390/su13116165. URL: https://www.mdpi.com/2071-1050/13/11/6165.

[45] Bart P. Knijnenburg et al. "Explaining the user experience of recommender systems". In: *User Modeling and User-Adapted Interaction* 22.4 (Oct. 2012), pp. 441–504. ISSN: 1573-1391. DOI: 10.1007/s11257-011-9118-4. URL: https://doi.org/10.1007/s11257-011-9118-4.

[46] Joseph A. Konstan and Loren G. Terveen. "Human-centered Recommender Systems: Origins, Advances, Challenges, and Opportunities". In: *AI Mag.* 42.3 (Sept. 2021), pp. 31–42. ISSN: 0738-4602. DOI: 10.1609/aimag.v42i3.18142. URL: https://doi.org/10.1609/aimag.v42i3.18142.

[47] Yehuda Koren. "Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '08. Las Vegas, Nevada, USA: Association for Computing Machinery, 2008, pp. 426–434. ISBN: 9781605581934. DOI: 10.1145/1401890.1401944. URL: https://doi.org/10.1145/1401890.1401944.

[48] Denis Kotkov et al. "The Tag Genome Dataset for Books". In: *Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22)*. 2022. DOI: 10.1145/3498366.3505833.

[49] Frank Lang. *Der Buechernarr*. [Online, accessed 24-November-2023]. URL: https://buechernarr.org/.

[50] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

[51] Juha Leino and Kari-Jouko Räihä. "Case Amazon: Ratings and Reviews as Part of Recommendations". In: *Proceedings of the 2007 ACM Conference on Recommender Systems*. RecSys '07. Minneapolis, MN, USA: Association for Computing Machinery, 2007, pp. 137–140. ISBN: 9781595937308. DOI: 10.1145/1297231.1297255. URL: https://doi.org/10.1145/1297231.1297255.

[52] Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. “Facing the cold start problem in recommender systems”. In: *Expert Systems with Applications* 41.4, Part 2 (2014), pp. 2065–2073. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2013.09.005. URL: https://www.sciencedirect.com/science/article/pii/S0957417413007240.

[53] G. Linden, B. Smith, and J. York. “Amazon.com recommendations: item-to-item collaborative filtering”. In: *IEEE Internet Computing* 7.1 (2003), pp. 76–80. DOI: 10.1109/MIC.2003.1167344.

[54] Minyu Liu. “Personalized Recommendation System Design for Library Resources through Deep Belief Networks”. In: *Mobile Information Systems* 2022 (July 2022), pp. 1–8. DOI: 10.1155/2022/7870724.

[55] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. “Content-based Recommender Systems: State of the Art and Trends”. In: *Recommender Systems Handbook*. Ed. by Francesco Ricci et al. Boston, MA: Springer US, 2011, pp. 73–105. ISBN: 978-0-387-85820-3. DOI: 10.1007/978-0-387-85820-3_3. URL: https://doi.org/10.1007/978-0-387-85820-3_3.

[56] Jie Lu et al. “Recommender system application developments: A survey”. In: *Decision Support Systems* 74 (2015), pp. 12–32. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2015.03.008. URL: https://www.sciencedirect.com/science/article/pii/S0167923615000627.

[57] Xin Luo et al. “An Efficient Non-Negative Matrix-Factorization-Based Approach to Collaborative Filtering for Recommender Systems”. In: *IEEE Transactions on Industrial Informatics* 10.2 (2014), pp. 1273–1284. DOI: 10.1109/TII.2014.2308433.

[58] Saranya Maneeroj and Atsuhiro Takasu. “Hybrid Recommender System Using Latent Features”. In: *2009 International Conference on Advanced Information Networking and Applications Workshops*. 2009, pp. 661–666. DOI: 10.1109/WAINA.2009.122.

[59] Praveena Mathew, Bincy Kuriakose, and Vinayak Hegde. “Book Recommendation System through content based and collaborative filtering method”. In: *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*. 2016, pp. 47–52. DOI: 10.1109/SAPIENCE.2016.7684166.

[60] Christian Matt et al. “Escaping from the filter bubble? The effects of novelty and serendipity on users’ evaluations of online recommendations”. In: (2014).

[61] Sean M. McNee, John Riedl, and Joseph A. Konstan. “Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems”. In: *CHI ’06 Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’06. Montréal, Québec, Canada: Association for Computing Machinery, 2006, pp. 1097–1101. ISBN: 1595932984. DOI: 10.1145/1125451.1125659. URL: https://doi.org/10.1145/1125451.1125659.

[62] Sean M. McNee, John Riedl, and Joseph A. Konstan. “Making Recommendations Better: An Analytic Model for Human-Recommender Interaction”. In: *CHI ’06 Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’06. Montréal, Québec, Canada: Association for Computing Machinery, 2006, pp. 1103–1108. ISBN: 1595932984. DOI: 10.1145/1125451.1125660. URL: https://doi.org/10.1145/1125451.1125660.

[63] *MovieLens. Non-commercial, personalized movie recommendations.* [Online, accessed 25-October-2023]. URL: https://movielens.org/.

[64] Tien T. Nguyen et al. “Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity”. In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW ’14. Seoul, Korea: Association for Computing Machinery, 2014, pp. 677–686. ISBN: 9781450327442. DOI: 10.1145/2566486.2568012. URL: https://doi.org/10.1145/2566486.2568012.

[65] Kalia Orphanou et al. “Mitigating Bias in Algorithmic Systems—A Fish-Eye View”. In: *ACM Comput. Surv.* 55.5 (Dec. 2022). ISSN: 0360-0300. DOI: 10.1145/3527152. URL: https://doi.org/10.1145/3527152.

[66] Michael J. Pazzani and Daniel Billsus. “Content-Based Recommendation Systems”. In: *The Adaptive Web: Methods and Strategies of Web Personalization*. Ed. by Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 325–341. ISBN: 978-3-540-72079-9. DOI: 10.1007/978-3-540-72079-9_10. URL: https://doi.org/10.1007/978-3-540-72079-9_10.

[67] Svenja Pieritz et al. “Personalised Recommendations in Mental Health Apps: The Impact of Autonomy and Data Sharing”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445523. URL: https://doi.org/10.1145/3411764.3445523.

[68] Ivens Portugal, Paulo Alencar, and Donald Cowan. "The use of machine learning algorithms in recommender systems: A systematic review". In: *Expert Systems with Applications* 97 (2018), pp. 205–227. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2017.12.020. URL: https://www.sciencedirect.com/science/article/pii/S0957417417308333.

[69] Dhanya Pramod and Prafulla Bafna. "Conversational recommender systems techniques, tools, acceptance, and adoption: A state of the art review". In: *Expert Systems with Applications* 203 (2022), p. 117539. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.117539. URL: https://www.sciencedirect.com/science/article/pii/S0957417422008612.

[70] Kiran R, Pradeep Kumar, and Bharat Bhasker. "DNNRec: A Novel Deep Learning Based Hybrid Recommender System". In: *Expert Syst. Appl.* 144.C (Apr. 2020). ISSN: 0957-4174. DOI: 10.1016/j.eswa.2019.113054. URL: https://doi.org/10.1016/j.eswa.2019.113054.

[71] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. URL: https://www.R-project.org/.

[72] *Reddit | Books*. [Online, accessed 24-November-2023]. URL: https://www.reddit.com/r/books/.

[73] *Reddit | printSF*. [Online, accessed 24-November-2023]. URL: https://www.reddit.com/r/printSF/.

[74] *Reddit | SciFI*. [Online, accessed 24-November-2023]. URL: https://www.reddit.com/r/scifi/.

[75] Steffen Rendle. "Factorization Machines". In: *2010 IEEE International Conference on Data Mining*. 2010, pp. 995–1000. DOI: 10.1109/ICDM.2010.127.

[76] Steffen Rendle et al. "Neural Collaborative Filtering vs. Matrix Factorization Revisited". In: *Proceedings of the 14th ACM Conference on Recommender Systems*. RecSys '20. Virtual Event, Brazil: Association for Computing Machinery, 2020, pp. 240–248. ISBN: 9781450375832. DOI: 10.1145/3383313.3412488. URL: https://doi.org/10.1145/3383313.3412488.

[77] Yves Rosseel. "lavaan: An R Package for Structural Equation Modeling". In: *Journal of Statistical Software* 48.2 (2012), pp. 1–36. DOI: 10.18637/jss.v048.i02.

[78] Ruslan Salakhutdinov and Andriy Mnih. "Bayesian Probabilistic Matrix Factorization Using Markov Chain Monte Carlo". In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. Helsinki, Finland: Association for Computing Machinery, 2008, pp. 880–887. ISBN: 9781605582054. DOI: 10.1145/1390156.1390267. URL: https://doi.org/10.1145/1390156.1390267.

[79] Badrul Sarwar et al. "Analysis of Recommendation Algorithms for E-Commerce". In: *Proceedings of the 2nd ACM Conference on Electronic Commerce*. EC '00. Minneapolis, Minnesota, USA: Association for Computing Machinery, 2000, pp. 158–167. ISBN: 1581132727. DOI: 10.1145/352871.352887. URL: https://doi.org/10.1145/352871.352887.

[80] Badrul Sarwar et al. "Item-Based Collaborative Filtering Recommendation Algorithms". In: *Proceedings of the 10th International Conference on World Wide Web*. WWW '01. Hong Kong, Hong Kong: Association for Computing Machinery, 2001, pp. 285–295. ISBN: 1581133480. DOI: 10.1145/371920.372071. URL: https://doi.org/10.1145/371920.372071.

[81] Markus Schedl et al. "Current challenges and visions in music recommender systems research". In: *International Journal of Multimedia Information Retrieval* 7.2 (Apr. 2018), pp. 95–116. DOI: 10.1007/s13735-018-0154-2. URL: https://doi.org/10.1007%2Fs13735-018-0154-2.

[82] Suvash Sedhain et al. "AutoRec: Autoencoders Meet Collaborative Filtering". In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15 Companion. Florence, Italy: Association for Computing Machinery, 2015, pp. 111–112. ISBN: 9781450334730. DOI: 10.1145/2740908.2742726. URL: https://doi.org/10.1145/2740908.2742726.

[83] Alexander Serenko, Nick Bontis, and Emily Hull. "Practical relevance of knowledge management and intellectual capital scholarly research: Books as knowledge translation agents". In: *Knowledge and Process Management* 18.1 (2011), pp. 1–9. DOI: https://doi.org/10.1002/kpm.363. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/kpm.363. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/kpm.363.

[84] Till Speicher et al. "Potential for Discrimination in Online Targeted Advertising". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning

Research. PMLR, Feb. 2018, pp. 5–19. URL: https://proceedings.mlr.press/v81/speicher18a.html.

[85] *Statista | Percentage of U.S. adults who use Reddit as of February 2021, by age group*. Details: United States; Pew Research Center; January 25 to February 8, 2021; 18 years and older; Telephone Interview [Online, accessed 26-Ocotber-2023]. URL: https://www.statista.com/statistics/261766/share-of-us-internet-users-who-use-reddit-by-age-group/.

[86] *Statista | U.S. Readers Are Getting Less Voracious*. [Online, accessed 26-Ocotber-2023]. URL: https://www.statista.com/chart/26572/average-number-of-books-read-by-us-residents-per-year/.

[87] Florian Strub, Romaric Gaudel, and Jérémie Mary. "Hybrid Recommender System Based on Autoencoders". In: *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*. DLRS 2016. Boston, MA, USA: Association for Computing Machinery, 2016, pp. 11–16. ISBN: 9781450347952. DOI: 10.1145/2988450.2988456. URL: https://doi.org/10.1145/2988450.2988456.

[88] Latanya Sweeney. "Discrimination in Online Ad Delivery". In: (2013). arXiv: 1301.6822 [cs.IR].

[89] Gábor Takács, István Pilászy, and Domonkos Tikk. "Applications of the Conjugate Gradient Method for Implicit Feedback Collaborative Filtering". In: *Proceedings of the Fifth ACM Conference on Recommender Systems*. RecSys '11. Chicago, Illinois, USA: Association for Computing Machinery, 2011, pp. 297–300. ISBN: 9781450306836. DOI: 10.1145/2043932.2043987. URL: https://doi.org/10.1145/2043932.2043987.

[90] Thiviyan Thanapalasingam et al. "Ontology-Based Recommendation of Editorial Products". In: *Lecture Notes in Computer Science*. Springer International Publishing, 2018, pp. 341–358. DOI: 10.1007/978-3-030-00668-6_21. URL: https://doi.org/10.1007%2F978-3-030-00668-6_21.

[91] Mike Thelwall and Kayvan Kousha. "Goodreads: A social network site for book readers". In: *Journal of the Association for Information Science and Technology* 68.4 (2017), pp. 972–983. DOI: https://doi.org/10.1002/asi.23733. eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23733. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23733.

[92] Yonghong Tian et al. "College Library Personalized Recommendation System Based on Hybrid Recommendation Algorithm". In: *Procedia CIRP* 83 (2019). 11th CIRP Conference on Industrial Product-Service Systems, pp. 490–494. ISSN: 2212-8271.

DOI: https://doi.org/10.1016/j.procir.2019.04.126. URL: https://www.sciencedirect.com/science/article/pii/S2212827119307401.

[93] Mengting Wan and Julian J. McAuley. “Item recommendation on monotonic behavior chains”. In: *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*. Ed. by Sole Pera et al. ACM, 2018, pp. 86–94. DOI: 10.1145/3240323.3240369. URL: https://doi.org/10.1145/3240323.3240369.

[94] Mengting Wan et al. “Fine-Grained Spoiler Detection from Large-Scale Review Corpora”. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, and Lluıs Màrquez. Association for Computational Linguistics, 2019, pp. 2605–2610. DOI: 10.18653/v1/p19-1248. URL: https://doi.org/10.18653/v1/p19-1248.

[95] Jun Wang, Arjen P. de Vries, and Marcel J. T. Reinders. “Unifying User-Based and Item-Based Collaborative Filtering Approaches by Similarity Fusion”. In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’06. Seattle, Washington, USA: Association for Computing Machinery, 2006, pp. 501–508. ISBN: 1595933697. DOI: 10.1145/1148170.1148257. URL: https://doi.org/10.1145/1148170.1148257.

[96] Martijn C. Willemsen, Mark P. Graus, and Bart P. Knijnenburg. “Understanding the role of latent feature diversification on choice difficulty and satisfaction”. In: *User Modeling and User-Adapted Interaction* 26.4 (Oct. 2016), pp. 347–389. ISSN: 1573-1391. DOI: 10.1007/s11257-016-9178-6. URL: https://doi.org/10.1007/s11257-016-9178-6.

[97] Hong Xie. “Shifts in information-seeking strategies in information retrieval in the digital age: Planned-situational model”. In: *Inf. Res.* 12 (2007). URL: https://api.semanticscholar.org/CorpusID:27864905.

[98] Feng Xue et al. “Deep Item-Based Collaborative Filtering for Top-N Recommendation”. In: *ACM Trans. Inf. Syst.* 37.3 (Apr. 2019). ISSN: 1046-8188. DOI: 10.1145/3314578. URL: https://doi.org/10.1145/3314578.

[99] Mi Zhang and Neil Hurley. “Avoiding monotony: Improving the diversity of recommendation lists”. In: Oct. 2008, pp. 123–130. DOI: 10.1145/1454008.1454030.

[100] Qian Zhang, Jie Lu, and Yaochu Jin. "Artificial intelligence in recommender systems". In: *Complex & Intelligent Systems* 7.1 (Feb. 2021), pp. 439–457. ISSN: 2198-6053. DOI: 10.1007/s40747-020-00212-w. URL: https://doi.org/10.1007/s40747-020-00212-w.

[101] Shuai Zhang et al. "Deep Learning Based Recommender System: A Survey and New Perspectives". In: *ACM Comput. Surv.* 52.1 (Feb. 2019). ISSN: 0360-0300. DOI: 10.1145/3285029. URL: https://doi.org/10.1145/3285029.

[102] Qian Zhao et al. "From Preference into Decision Making: Modeling User Interactions in Recommender Systems". In: *Proceedings of the 13th ACM Conference on Recommender Systems*. RecSys '19. Copenhagen, Denmark: Association for Computing Machinery, 2019, pp. 29–33. ISBN: 9781450362436. DOI: 10.1145/3298689.3347065. URL: https://doi.org/10.1145/3298689.3347065.

[103] Qian Zhao et al. "Gaze Prediction for Recommender Systems". In: *Proceedings of the 10th ACM Conference on Recommender Systems*. RecSys '16. Boston, Massachusetts, USA: Association for Computing Machinery, 2016, pp. 131–138. ISBN: 9781450340359. DOI: 10.1145/2959100.2959150. URL: https://doi.org/10.1145/2959100.2959150.

[104] Qian Zhao et al. "Interpreting User Inaction in Recommender Systems". In: *Proceedings of the 12th ACM Conference on Recommender Systems*. RecSys '18. Vancouver, British Columbia, Canada: Association for Computing Machinery, 2018, pp. 40–48. ISBN: 9781450359016. DOI: 10.1145/3240323.3240366. URL: https://doi.org/10.1145/3240323.3240366.

[105] Tao Zhou et al. "Solving the apparent diversity-accuracy dilemma of recommender systems". In: *Proceedings of the National Academy of Sciences* 107.10 (Feb. 2010), pp. 4511–4515. DOI: 10.1073/pnas.1000488107. URL: https://doi.org/10.1073%2Fpnas.1000488107.

[106] Yunhong Zhou et al. "Large-Scale Parallel Collaborative Filtering for the Netflix Prize". In: *Algorithmic Aspects in Information and Management*. Ed. by Rudolf Fleischer and Jinhui Xu. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 337–348. ISBN: 978-3-540-68880-8.

[107] Cai-Nicolas Ziegler et al. "Improving Recommendation Lists through Topic Diversification". In: *Proceedings of the 14th International Conference on World Wide Web*. WWW '05. Chiba, Japan: Association for Computing Machinery, 2005, pp. 22–32.

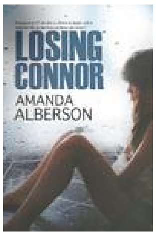
ISBN: 1595930469. DOI: 10.1145/1060745.1060754. URL: https://doi.org/10.1145/1060745.1060754.

Appendix A

Complete Recommendation Lists

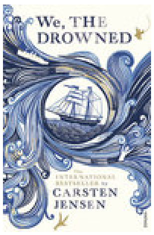
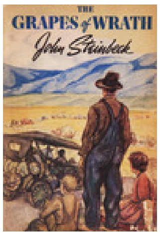
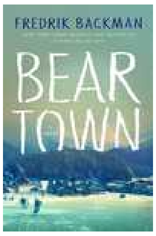
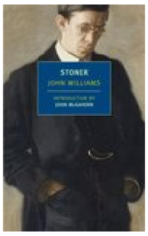
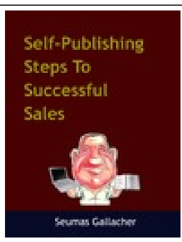
This appendix shows the full recommendation lists generated by each algorithm the users were presented with in the study.

Table A.1: Recommendation lists generated by **Item-Item** algorithm

Rank	Cover	Title	Year	Publisher	Rating
1		Losing Connor	2012	Amanda Alberson	4.51
2		Free to Be...You and Me	2008	Running Press	4.39
3		A Shot at the Big Time	2016	Christina McMullen	4.79

Continued on next page

Table A.1: Recommendation lists generated by **Item-Item** algorithm (Continued)

4		We, the Drowned	2011	Vintagen	4.22
5		The Grapes of Wrath	1939	Viking Press (NY)	3.93
6		Beartown	2017	Atria Books	4.31
7		Stoner	2003	NYRB Classics	4.28
8		Self-Publishing Steps To Successful Sales	2014	SGC Publishing	4.64

Continued on next page

Table A.1: Recommendation lists generated by **Item-Item** algorithm (Continued)

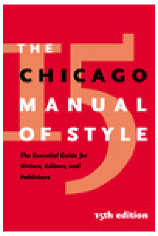
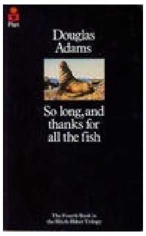
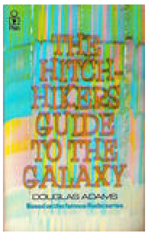
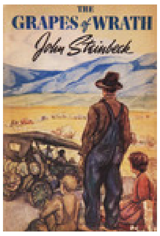
9		Strangers in Paradise	1999	Abstract Studio	4.38
10		Chicago Manual of Style	2003	University of Chicago Press	4.26

Table A.2: Recommendation lists generated by **User-User** algorithm

Rank	Cover	Title	Year	Publisher	Rating
1		So Long, and Thanks for All the Fish	1985	Pan	4.08
2		The Doors of Time	2009	Self Published on LJ	4.2

Continued on next page

Table A.2: Recommendation lists generated by **User-User** algorithm (Continued)

3		The Hitchhiker's Guide to the Galaxy	1979	Pan Books	4.2
4		The Duchess War	2013	Courtney Milan	3.82
5		At Least There's The Football	2012	Self Published	4.43
6		The Grapes of Wrath	1939	Viking Press (NY)	3.93
7		Act Like It	2015	Carina Press	3.92

Continued on next page

Table A.2: Recommendation lists generated by **User-User** algorithm (Continued)

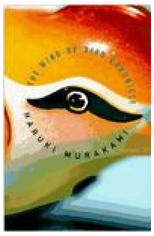
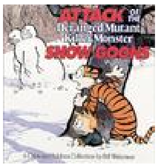
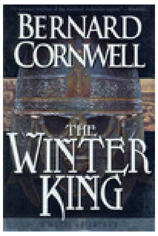
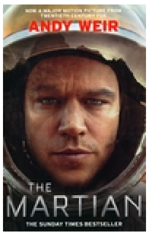
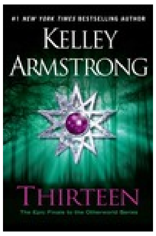
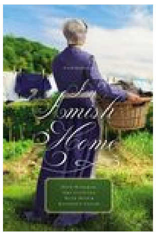
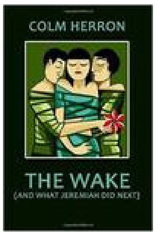
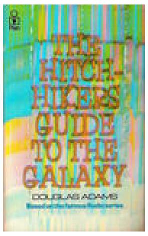
8		The Wind-Up Bird Chronicle	1997	Alfred A. Knopf	4.17
9		Attack of the Deranged Mutant Killer Monster Snow Goons	1992	Turtleback Books	4.72
10		The Winter King	1997	St. Martin's Griffin	4.27

Table A.3: Recommendation lists generated by **ALS** algorithm

Rank	**Cover**	**Title**	**Year**	**Publisher**	**Rating**
1		The Martian	2015	Del Rey	4.39

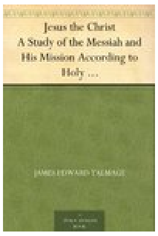
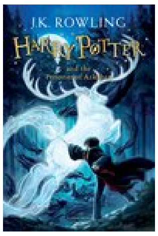
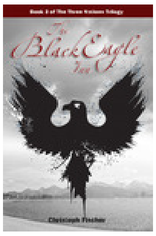
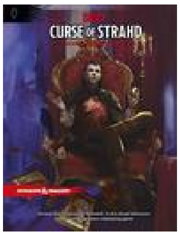
Continued on next page

Table A.3: Recommendation lists generated by **ALS** algorithm (Continued)

2		Thirteen	2012	Penguin Group	4.28
3		Ready Player One	2012	Broadway Books	4.31
4		An Amish Home	2017	Thomas Nelson	4.49
5		The Wake	2014	Nuascealta	4.42
6		The Hitchhiker's Guide to the Galaxy	1979	Pan Books	4.2

Continued on next page

Table A.3: Recommendation lists generated by **ALS** algorithm (Continued)

7		Jesus the Christ	1915	James E. Talmage	4.63
8		Harry Potter and the Prisoner of Azkaban	2014	Bloomsbury Childrens Books	4.53
9		The Black Eagle Inn	2013	Turtleback Books	4.51
10		Curse of Strahd	2016	Wizards of the Coast	4.5

Appendix B

Full Lavaan Ouput

This appendix contains the full lavaan output for both SEM and CFA from chapter 4.

B.1 Confirmatory Factor Analysis

```
model <- '
  Acc =~ NA * AccAppealing + AccBest + AccBad + AccAtTop
  Div =~ NA * DivSimilar + DivVaried + DivMoods + DivTastes
  Und =~ NA * UndTaste + UndTrust + UndPersonalized + UndMainstream
  Sat =~ NA * SatFind + SatRecommend + SatValuable + SatMobile + SatSat
  Nov =~ NA * NovUnexpected + NovFamiliar + NovSurprising + NovUnthought +
      NovFewerNew
  Acc ~~ 1*Acc
  Div ~~ 1*Div
  Nov ~~ 1*Nov
  Sat ~~ 1*Sat
  Und ~~ 1*Und
'
fit <- sem(model, data = combined_df)

> summary(fit, fit.measures = FALSE)

lavaan 0.6.16 ended normally after 89 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        54
```

```
  Number of observations                            97

Model Test User Model:

  Test statistic                               401.659
  Degrees of freedom                               199
  P-value (Chi-square)                           0.000

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  Acc =~
    AccAppealing      0.925    0.077   12.068    0.000
    AccBest           0.899    0.078   11.504    0.000
    AccBad           -0.720    0.088   -8.228    0.000
    AccAtTop          0.778    0.085    9.176    0.000
  Div =~
    DivSimilar        0.190    0.112    1.688    0.091
    DivVaried        -0.703    0.103   -6.830    0.000
    DivMoods         -0.814    0.103   -7.924    0.000
    DivTastes        -0.619    0.104   -5.977    0.000
  Und =~
    UndTaste          0.925    0.076   12.108    0.000
    UndTrust          0.914    0.077   11.876    0.000
    UndPersonalizd    0.909    0.077   11.762    0.000
    UndMainstream     0.036    0.101    0.355    0.722
  Sat =~
    SatFind           0.892    0.078   11.373    0.000
    SatRecommend      0.862    0.080   10.738    0.000
    SatValuable       0.880    0.079   11.111    0.000
    SatMobile         0.885    0.079   11.216    0.000
    SatSat            0.917    0.077   11.929    0.000
  Nov =~
    NovUnexpected     0.490    0.096    5.112    0.000
```

```
    NovFamiliar     -0.841    0.085   -9.942    0.000
    NovSurprising   -0.528    0.095   -5.554    0.000
    NovUnthought     0.518    0.095    5.439    0.000
    NovFewerNew     -0.029    0.101   -0.285    0.775

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
  Acc ~~
    Div             -0.146    0.117   -1.251    0.211
    Und              1.001    0.011   93.466    0.000
    Sat              0.971    0.014   67.951    0.000
    Nov             -0.980    0.035  -28.153    0.000
  Div ~~
    Und             -0.042    0.118   -0.357    0.721
    Sat             -0.076    0.117   -0.649    0.516
    Nov             -0.033    0.129   -0.259    0.796
  Und ~~
    Sat              0.998    0.009  112.637    0.000
    Nov             -1.001    0.032  -31.218    0.000
  Sat ~~
    Nov             -1.005    0.031  -32.353    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
    Acc              1.000
    Div              1.000
    Nov              1.000
    Sat              1.000
    Und              1.000
   .AccAppealing     0.135    0.025    5.417    0.000
   .AccBest          0.181    0.030    5.968    0.000
   .AccBad           0.471    0.070    6.758    0.000
   .AccAtTop         0.384    0.058    6.661    0.000
   .DivSimilar       0.954    0.138    6.895    0.000
   .DivVaried        0.495    0.104    4.744    0.000
   .DivMoods         0.327    0.110    2.966    0.003
   .DivTastes        0.607    0.107    5.678    0.000
   .UndTaste         0.135    0.022    6.077    0.000
   .UndTrust         0.154    0.025    6.249    0.000
   .UndPersonalizd   0.163    0.026    6.316    0.000
```

```
  .UndMainstream    0.988    0.142    6.964    0.000
  .SatFind          0.193    0.031    6.237    0.000
  .SatRecommend     0.247    0.038    6.442    0.000
  .SatValuable      0.215    0.034    6.334    0.000
  .SatMobile        0.206    0.033    6.298    0.000
  .SatSat           0.148    0.025    5.934    0.000
  .NovUnexpected    0.749    0.109    6.898    0.000
  .NovFamiliar      0.282    0.058    4.875    0.000
  .NovSurprising    0.711    0.104    6.869    0.000
  .NovUnthought     0.721    0.105    6.877    0.000
  .NovFewerNew      0.989    0.142    6.964    0.000
```

B.2 Structural Equation Modeling

```
model <- '
  Div =~ NA * DivVaried + DivMoods + DivTastes
  Acc =~ NA * AccAppealing + AccBest + AccBad + AccAtTop
  Nov =~ NA * NovUnexpected + NovFamiliar + NovSurprising + NovUnthought
  Div ~~ 1*Div
  Acc ~~ 1*Acc
  Nov ~~ 1*Nov

  # Regressions
  CondALSUU ~ CondIIUU
  Acc ~ Div + CondALSUU
  Nov ~ CondIIUU
  FirstImpression ~ Acc + Nov
'

fit <- sem(model, data = combined_df)

> summary(fit, fit.measures = FALSE)

lavaan 0.6.16 ended normally after 36 iterations

Estimator                                         ML
  Optimization method                         NLMINB
  Number of model parameters                      30
```

```
  Number of observations                            97

Model Test User Model:

  Test statistic                                240.174
  Degrees of freedom                                 74
  P-value (Chi-square)                            0.000

Parameter Estimates:

  Standard errors                              Standard
  Information                                  Expected
  Information saturated (h1) model           Structured

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  Div =~
    DivVaried         0.657    0.106    6.212    0.000
    DivMoods          0.863    0.109    7.953    0.000
    DivTastes         0.609    0.105    5.782    0.000
  Acc =~
    AccAppealing      0.713    0.064   11.133    0.000
    AccBest           0.731    0.063   11.593    0.000
    AccBad           -0.567    0.071   -7.974    0.000
    AccAtTop          0.616    0.069    8.946    0.000
  Nov =~
    NovUnexpected     0.671    0.102    6.568    0.000
    NovFamiliar      -0.600    0.103   -5.827    0.000
    NovSurprising    -0.461    0.106   -4.335    0.000
    NovUnthought      0.698    0.102    6.852    0.000

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  CondALSUU ~
    CondIIUU         -0.565    0.084   -6.736    0.000
  Acc ~
    Div               0.234    0.121    1.928    0.054
    CondALSUU        -1.533    0.252   -6.083    0.000
  Nov ~
```

```
    CondIIUU          -0.536    0.251   -2.135    0.033
  FirstImpression ~
    Acc                0.665    0.058   11.411    0.000
    Nov               -0.136    0.048   -2.835    0.005

Variances:
                   Estimate Std.Err z-value P(>|z|)
    Div               1.000
   .Acc               1.000
   .Nov               1.000
   .DivVaried         0.558    0.109    5.100    0.000
   .DivMoods          0.244    0.132    1.851    0.064
   .DivTastes         0.619    0.110    5.636    0.000
   .AccAppealing      0.176    0.033    5.265    0.000
   .AccBest           0.134    0.029    4.604    0.000
   .AccBad            0.475    0.072    6.581    0.000
   .AccAtTop          0.381    0.060    6.396    0.000
   .NovUnexpected     0.510    0.107    4.748    0.000
   .NovFamiliar       0.606    0.110    5.498    0.000
   .NovSurprising     0.763    0.121    6.296    0.000
   .NovUnthought      0.470    0.107    4.371    0.000
   .CondALSUU         0.157    0.023    6.964    0.000
   .FirstImpressin    0.122    0.026    4.669    0.000
```

B.3 Structural Equation Modeling with Demographic Data

```
model <- '
  Div =~ NA * DivVaried + DivMoods + DivTastes
  Acc =~ NA * AccAppealing + AccBest + AccBad + AccAtTop
  Nov =~ NA * NovUnexpected + NovFamiliar + NovSurprising + NovUnthought
  Div ~~ 1*Div
  Acc ~~ 1*Acc
  Nov ~~ 1*Nov

  # Regressions
  CondALSUU ~ CondIIUU
  Acc ~ Div + CondALSUU + expert_score
  Nov ~ CondIIUU
```

```
  FirstImpression ~ Acc + Nov
'

fit <- sem(model, data = combined_df)

> summary(fit, fit.measures = FALSE)

lavaan 0.6.16 ended normally after 36 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        30

  Number of observations                            97

Model Test User Model:

  Test statistic                               240.174
  Degrees of freedom                                74
  P-value (Chi-square)                           0.000

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  Div =~
    DivVaried         0.657    0.106    6.212    0.000
    DivMoods          0.863    0.109    7.953    0.000
    DivTastes         0.609    0.105    5.782    0.000
  Acc =~
    AccAppealing      0.713    0.064   11.133    0.000
    AccBest           0.731    0.063   11.593    0.000
    AccBad           -0.567    0.071   -7.974    0.000
    AccAtTop          0.616    0.069    8.946    0.000
  Nov =~
    NovUnexpected     0.671    0.102    6.568    0.000
```

```
    NovFamiliar    -0.600    0.103   -5.827    0.000
    NovSurprising  -0.461    0.106   -4.335    0.000
    NovUnthought    0.698    0.102    6.852    0.000

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  CondALSUU ~
    CondIIUU         -0.565    0.084   -6.736    0.000
  Acc ~
    Div               0.234    0.121    1.928    0.054
    CondALSUU        -1.533    0.252   -6.083    0.000
  Nov ~
    CondIIUU         -0.536    0.251   -2.135    0.033
  FirstImpression ~
    Acc               0.665    0.058   11.411    0.000
    Nov              -0.136    0.048   -2.835    0.005

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
    Div               1.000
   .Acc               1.000
   .Nov               1.000
   .DivVaried         0.558    0.109    5.100    0.000
   .DivMoods          0.244    0.132    1.851    0.064
   .DivTastes         0.619    0.110    5.636    0.000
   .AccAppealing      0.176    0.033    5.265    0.000
   .AccBest           0.134    0.029    4.604    0.000
   .AccBad            0.475    0.072    6.581    0.000
   .AccAtTop          0.381    0.060    6.396    0.000
   .NovUnexpected     0.510    0.107    4.748    0.000
   .NovFamiliar       0.606    0.110    5.498    0.000
   .NovSurprising     0.763    0.121    6.296    0.000
   .NovUnthought      0.470    0.107    4.371    0.000
   .CondALSUU         0.157    0.023    6.964    0.000
   .FirstImpressin    0.122    0.026    4.669    0.000
```