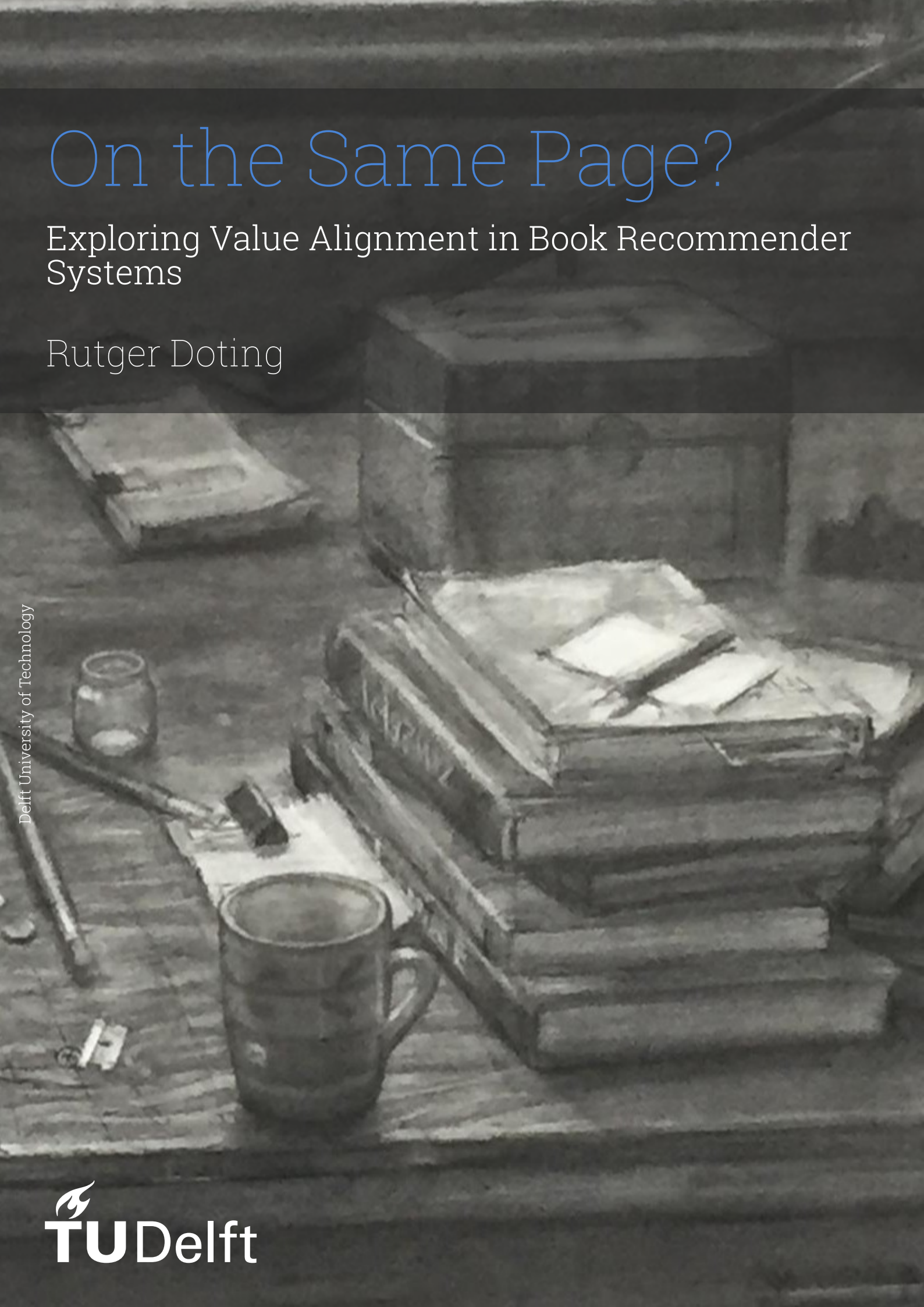
On the Same Page?

Exploring Value Alignment in Book Recommender Systems

Rutger Doting



Delft University of Technology

TU Delft

On the Same Page?

Exploring Value Alignment in Book Recommender Systems

by

Rutger Doting

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday June 15, 2026 at 13:00.

Student number: 5633427
Project duration: June 2, 2025 – June 15, 2026
Thesis committee: Dr. M. S. Pera, TU Delft, supervisor
Dr. L. Cavalcante Siebert, TU Delft

Cover: Sjoerd Doting, Studio Table and Window, pencil on paper, detail (https://sjoerddoting.com)

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Acknowledgments

I would like to express my sincere gratitude to my supervisors, Sole Pera and Garrett Allen, whose guidance, expertise, and encouragement were invaluable throughout this project.

I would also like to thank Stefan Buijsman, Bastiaan Hoorneman, and Doris Bukman for their time, input, and support at various stages of this thesis.

Lastly, I would like to thank my family and my girlfriend for their extensive support during this whole process, and their willingness to proofread this thesis whenever I asked them to.

Rutger Doting
Delft, June 2026

Summary

Recommender systems (RSs) have become a central part of daily digital life, shaping which items, people, and opportunities users are exposed to at scale. Given that personal values are fundamental to individual identity, decision-making, and consumer behavior, and that RSs learn from interaction data that is itself shaped by these values, the question arises whether the recommendations produced by standard RSs already reflect users' personal values. To date, no prior work has empirically investigated this question.

This thesis addresses that gap by examining the extent to which user values are reflected in recommendation outcomes, and whether explicitly incorporating value information can improve this alignment. Using Schwartz's Theory of Basic Human Values as a theoretical framework, we conduct an offline experiment on the Goodreads dataset. We construct value profiles for both users and recommended items using the Personal Values Dictionary, which maps over a thousand English words to their corresponding Schwartz value. These profiles are derived from user reviews and book descriptions respectively, and are used to measure the alignment between a user's personal values and the values embedded in their recommendations.

Our results show that standard RSs exhibit a weak but positive degree of value alignment, suggesting that interaction-based optimization procedures partially capture users' values without explicitly modeling them. Furthermore, we find that explicitly incorporating user value profiles as features within the RS increases this alignment. These findings carry important implications for the design of value-aware recommender systems, and suggest that early integration of value information is a promising direction for future research.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

Abbreviation	Definition
CARS	Context-Aware Recommender System
CF	Collaborative Filtering
CBF	Content-Based Filtering
ILD	Intra-list Diversity
MF	Matrix Factorization
MRR	Mean Reciprocal Rank
NDCG	Normalized Discounted Cumulative Gain
PVD	Personal Values Dictionary
RLVP	Recommendation List Value Profile
RS	Recommender System
SRS	Sequential Recommender System
SVS	Schwartz Value Survey
PVQ	Portrait Values Questionnaire
UVP	User Value Profile

1

Introduction

Recommender Systems (RSs) are intelligent software tools that leverage historical data, contextual signals, and predictive modeling techniques to identify and prioritize items that are likely to be of interest to specific users [178]. They are designed to mitigate information and choice overload by personalizing content selection, filtering, and presentation.

These systems have become ubiquitous in the digital landscape, with their use vastly increasing in recent years [149]. They are considered an integral part of daily life and personal routines, to the point where it is now difficult to find an online service that does not employ some form of recommendation algorithm [65]. Operating in the background of interactions across many of the world's largest platforms and apps, these systems select, filter, and personalize content [199]. By doing so, they help users find entertainment [83, 92], romantic partners [209], news [49], products [216], holiday destinations [73], job opportunities [52], and more. In this manner, these RSs shape which items, people, and opportunities users are exposed to at scale. Given that such exposure directly influences an individual's decisions and lifestyle [132], recent research has stressed the need for aligning these systems with human values to minimize their negative impacts [115, 97]. Consequently, understanding the relationship between RSs and the values of the individuals they serve becomes a natural and pressing concern.

Values are among the most fundamental factors shaping human life, defining what individuals and groups consider important and guiding their actions and decisions [37]. Following value-sensitive design literature, we define values as the things that individuals or groups consider to be important in life [69]. Values determine the ideal version of how one should live [117]. By doing so, values are the essential constituents of human well-being [133], and serve as the core of personhood [90]. In this capacity, values influence and motivate the actions of an individual [61, 185]. They play an important role in decision making and behavioral diversity [221], and impact consumer behavior [223, 130]. RSs, in turn, learn from this consumption data to generate recommendations. This raises the question: to what extent are the values of individual users reflected in the recommendations that RSs produce?

Given how central values are to individual identity and behavior, there are numerous ways in which they may be reflected in recommender systems. As the core of personhood, values influence and motivate individual action [61, 185], play an important role in decision-making and behavioral diversity [221], and shape consumer behavior [223, 130]. Thus, the values an individual holds will influence what they consume. Since RSs are built upon interaction data derived from this consumption behavior, the interactions that drive these systems are themselves (partially) an result of an individual's values. This raises the possibility that such values may, in turn, influence the functioning of the RS.

Another way in which values might influence RSs is through the content that individuals interact with. Prior research has identified strong correlations between specific value dimensions and content preferences [22, 117]. This means that when a standard system aligns a recommendation with a user's past ratings, it could inadvertently align these recommendations with the values that motivated those ratings. Furthermore, because values are acquired through socialization and shared experience, users tend to form behavioral clusters based on shared principles [82]. As a result, standard RS techniques could use

these clusters to surface items that resonate with the collective values of similar user groups [195]. Finally, prior work has found that items that align with their values are more appealing to individuals [225]. Thus, users might interact more with items that are aligned with their values, and this might be noticed by RSs.

Despite these myriad possible influences that values can have on RSs, to the best of our knowledge, no prior work has investigated what values are present in the recommendations. Because of this, it is unclear whether standard interaction-based optimization procedures may already capture users' values, and reflect them in the recommendations. This creates an empirical gap in the literature: do recommendations generated by traditional recommender systems already exhibit alignment with users' values?

To address this gap, we investigate how user values are reflected in recommendation outcomes and whether explicitly incorporating value information can improve alignment. Stated formally, we investigate the following research questions:

- **RQ1:** To what extent do recommendations generated by a standard recommender system align with users' personal values?
- **RQ2:** Does incorporating user values as features within a recommender increase this alignment?

To empirically investigate value alignment in recommender systems, we must first operationalize the concept of personal values using a robust and validated framework. Among existing frameworks, Schwartz's Theory of Basic Human Values [185] offers the most extensively validated and cross-culturally tested model. This theory defines values as concepts or beliefs about desirable end states that guide the evaluation of events and people [146]. Based on samples from 20 countries, Schwartz identified 10 value types, which are both comprehensive (they cover all the types of values that people attach importance to) and universal (they are present in all cultures). These values are organized according to their relative importance for each individual in a value hierarchy, and differences between individuals can be characterized by differences in these value hierarchies [185, 91].

Using this operationalization of values, we conduct an offline experiment using the Goodreads dataset [226, 227], which consists of data scraped from the book review site `goodreads.com` in late 2017. Conducting an offline experiment allows us to systematically compare baseline and value-aware recommendations while controlling for confounding variables. We first establish a baseline for value alignment in recommender systems by generating recommendations without incorporating value-based user features. We then create a value profile for the user and recommendations using the Personal Values Dictionary, introduced by Ponizovskiy et al. [169]. This validated dictionary links 1080 English words with their corresponding Schwartz value. Through the use of this dictionary, we construct value profiles based on free-text reviews (user value profile) and book descriptions (recommendation list value profile). We subsequently compare the alignment of the user value profiles and the recommendation list value profiles. Building upon this baseline, we then integrate user value profiles as explicit features within the RS, generate a new set of recommendations, and reassess the degree of value alignment.

By conducting this experiment, this thesis makes the following contributions:

- **Empirical insight into value alignment in recommender systems:** This thesis provides an empirical investigation into the current state of value alignment in recommender systems by conducting an offline experiment.
- **Exploratory analysis of value-aware recommendations:** Through a comparative analysis of baseline and value-aware recommendations, this work explores how explicitly incorporating user value profiles influences the alignment between recommendations and users' personal values.

Together, these contributions advance understanding of the role personal values play in personalized recommendations.

2

Background

In this chapter, we present background information regarding recommenders systems and Schwartz's Theory of Basic Human Values, as these are the central elements of our exploration.

2.1. Recommender Systems

RSs are software designed to identify items (e.g., books, movies, songs, or news articles) likely to be of interest to specific users based on historical data, contextual signals, and predictive modeling techniques [178]. RSs play a critical role in mitigating information overload in large-scale digital environments such as e-commerce platforms and media streaming services, where users must navigate extensive catalogs of available content. By prioritizing relevant items, they reduce perceived effort and improve user engagement and satisfaction [122]. RSs are broadly categorized into non-personalized and personalized approaches.

Non-personalized recommenders operate independently of individual user preferences. An example is the "Most Popular" recommender, which provides recommendations based on aggregate interaction statistics such as purchase frequency or view counts [7]. Non-personalized methods are particularly useful in settings where user-level data are unavailable.

While non-personalized recommenders serve as useful baselines in certain contexts, the majority of RS research, focuses on *personalized* recommenders [178]. These recommenders generate user-specific recommendations by modeling individual preferences. Formally, the personalized recommendation task can be expressed as estimating a relevance function [5], shown in Equation (2.1).

$$f(u, i) = \hat{r}_{ui} \tag{2.1}$$

where $\hat{r}_{ui}$ denotes the predicted utility of item i for user u:

For this recommendation task, RSs typically use data associated with one of three types of objects: users, items, and interactions of users with these items [178]. Though information such as user attributes (e.g., age, gender, personality) and item features (e.g., genre, price, size) can enrich RS models, their acquisition is often impractical. User data is frequently unavailable or difficult to obtain, and item features may require advanced techniques such as natural language processing or image analysis to extract. Consequently, the vast majority of RSs rely primarily on interaction data to infer user preferences [7, 178], and any auxiliary data available to a RS beyond this data is referred to as "side information" [244].

The interaction data that RSs utilize can take many forms, such as clicking, browsing, buying, or user feedback. User feedback plays an important part in recommender systems research, since it is used to determine user preferences [4]. User feedback can be divided into explicit and implicit feedback. Explicit feedback consists of user-provided signals such as ratings, likes, or textual reviews. These signals encode clear preference intensity and are typically of high quality [178]. However, users rarely

provide explicit feedback, leading to sparsity issues [94]. Therefore, implicit feedback has become the main type of user feedback used in RSs [247]. Implicit feedback is inferred from user behavior, including clicks, dwell time, browsing activity, and purchase history. Although abundant, implicit signals are noisy and ambiguous [247], as interactions do not necessarily imply positive preference [94].

Different methods have been developed that use these data sources for the recommendation task. These methods can be categorized into three general approaches based on how recommendations are made: content-based, collaborative filtering, and hybrid [5, 30, 178]. A visual overview of the different methods is shown in Figure 2.1, which is inspired by the classification of Ping et al. [168]; we refer readers to Ricci et al. [178] for an excellent in-depth overview of these approaches. While we acknowledge other recommendation approaches (e.g., content-based filtering) for completeness, this work focuses on collaborative filtering and thus discusses it in depth.

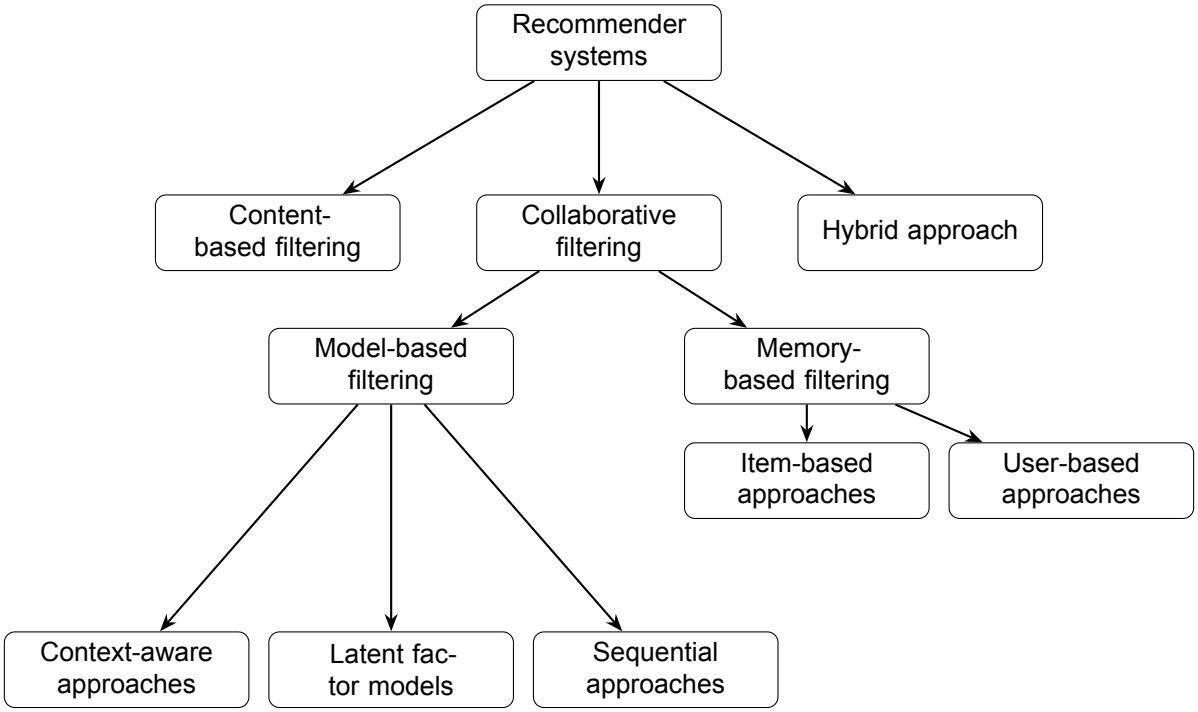


Figure 2.1: An overview of the different types of recommender systems, categorized by algorithmic logic and model representation. For model-based filtering, we only specify the approaches used in this thesis, due to space constraints.

2.1.1. Content-Based Filtering

Content-Based Filtering (CBF) recommenders suggest items whose attributes are similar to those previously preferred by the user [163]. As the name implies, these methods focus on the content of the items. The system computes the similarity between candidate items and items the user previously interacted with using feature representations derived from metadata (e.g., genre, textual descriptions, tags). This reliance on the item's attributes gives this method the advantages of interpretability and independence from other users' data.

The use of item attributes for recommendation also highlights a drawback of CBF: since performance depends on the availability and richness of item features, limited content analysis is a major constraint of this method. When content analysis is limited, the resulting mathematical representation of items is suboptimal, which leads to inaccurate recommendations [95]. Furthermore, if no descriptive features are available at all for an item, a content-based system cannot function or provide any suggestions [154].

Another limitation is the risk of overspecialization. Overspecialization occurs because CBF is designed to find items as similar as possible to those a user has liked in the past. Consequently, these systems tend to provide "obvious" recommendations because they rely on matching specific keywords or content features. Besides reducing the amount of novelty of these recommendations, this also limits the amount

of serendipity, or the “lucky discovery” of surprisingly relevant items [74].

2.1.2. Collaborative Filtering

Collaborative Filtering (CF) operates under the assumption that users with similar historical preferences will exhibit similar future preferences [183] and uses collective behavioral patterns to exploit this. Because these methods produce recommendations based on usage patterns, there is no need for external information about either items or users. Due to this flexibility, CF has become the most popular method in RS [178]. Below, we discuss the CF methods used in this thesis.

Neighborhood-Based Methods

Neighborhood-based methods use the relationships between users or items to suggest items. These methods are typically categorized into item-based and user-based approaches. User-based approaches predict item relevance by aggregating ratings from users with similar preferences, while item-based approaches use similarity among items previously rated by the target user. Similarity is typically computed using cosine similarity, Pearson correlation, or adjusted cosine metrics [183].

Latent Factor Models

Latent factor models learn low-dimensional embeddings for users and items using techniques such as matrix factorization [124]. These techniques decompose the user-item interaction matrix into two lower-dimensional matrices, a user matrix and an item matrix, in a process called factorization [11]. The resulting matrices are called *latent factors*. “Factor” refers to the factorization process, and “latent” refers to the fact that these factors are not directly observable in the data [142]. Instead, they are learned during training to capture hidden patterns, such as user preferences (e.g., a user’s affinity for a specific theme) and item attributes (e.g., a movie’s underlying themes) [7]. Factors can be seen as axes of variation that define a shared, low-dimensional space where both users and items are represented [7]. These axes capture the underlying patterns and essential structure of the interaction data. These methods use the user latent factor p_u and the item latent vector q_i for preference prediction via inner products in latent space, such as in Equation (2.2).

$$\hat{r}_{ui} = p_u^\top q_i \tag{2.2}$$

Context-Aware Recommender Systems

context-aware recommender systems (CARSs) are based on the insight that the context of recommendation is important in many applications, and incorporating this information can increase the quality of the recommendations [3]. These systems differ from traditional RSs by incorporating *contextual information*. This term refers to any additional information that affects a user’s preferences besides user and item characteristics [222]. These RSs include context in the recommendation task, therefore changing the recommendation task expressed in Equation (2.1) into the equation seen in Equation (2.3) [3].

$$f(u, i, c) = \hat{r}_{ui} \tag{2.3}$$

In addition to user information (u) and item information (i), this function takes contextual information (c) as input. Factors that are often included as contextual information are: time [123], location [8], and social information [195]. This last factor, social information, allows for a particularly wide area of use cases. For example, this information can be used to recommend missing nodes in a network or recommend products using social cues [7].

Sequential Recommender Systems

Many traditional recommendation algorithms are time-agnostic and ignore the chronological order of user interactions [101]. These methods assume that every past user interaction carries equal weight in determining their current preferences [230]. Breaking from this time-agnostic approach, sequential recommender systems (SRSs) assume that the temporal order of actions is often as important as the actions themselves [7]. For example, it makes sense to recommend coffee pods after a user buys a coffee machine, but not the other way around. Therefore, SRSs view the user-item interaction data as an interrelated sequence over time, instead of a series of isolated interactions [251]. A sequence refers to an ordered set of objects [171].

SRSs are a specialized form of sequence-aware RSs, which refers to any RS that considers information from sequentially ordered user-item interaction logs [171]. SRSs can be distinguished from the other main group in sequence-aware RSs, session-based RSs, by the fact that in SRSs longer term user histories are available, while session-based RSs rely on interaction sequence in a single session [101]. Based on this interaction sequence, SRSs attempt to predict the next item in the sequence [242]. This identification of sequential patterns is not used in traditional matrix completion methods such as collaborative filtering [171]. This method can be formalized in the following way, based on Zhou et al. [251]:

Let $u \in \mathcal{U}$ and $i \in \mathcal{I}$ denote a user and an item, where $\mathcal{U}$ and $\mathcal{I}$ are the sets of all users and items, respectively. For each user u, we define an ordered interaction history as a tuple of n interactions:

$$S_u = \langle s_u^1, s_u^2, \ldots, s_u^n \rangle, \tag{2.4}$$

where $n = |S_u|$ denotes the number of observed interactions for user u. Each interaction at step k is represented as a structured record:

$$s_u^k = \langle i_u^k,\, t_u^k,\, b_u^k,\, \ldots \rangle, \tag{2.5}$$

where i_u^k is the item interacted with by user u at position k; t_u^k is the timestamp of the interaction; b_u^k is the interaction type; and $\ldots$ denotes optional contextual information.

The primary objective of SRSs is to learn the underlying temporal dependencies and behavioral patterns embedded in S_u, in order to predict future user-item interactions [104]. Based on this objective, the outcomes can be presented to the user in multiple ways, such as a ranked list or a sequence of items [230].

Challenges

The specific nature of CF, with the reliance on users' behavioral patterns, brings its own challenges. Three challenges are most prominent for CF: the cold start problem, data sparsity, and scalability.

The cold start problem occurs when the system lacks sufficient information to make accurate predictions for new users or new items [131]. For example, when a new user joins, they do not have any interaction data, which makes it impossible to calculate similarity with other users. Similarly, in item-based CF, the new items do not have any interactions, so similarity to other items cannot be calculated.

Data sparsity arises because most users interact with only a tiny fraction of the available item catalog, leaving the user-item interaction matrix mostly empty. This makes similarity measures between users unreliable, because two users have few or no ratings in common. It also leads to neighbor transitivity, where similar users cannot be connected because they have insufficient common interactions [10].

Another important problem for CF is scalability. Large interaction datasets increase computational cost, because the cost of making recommendations often scales linearly with the data [7]. This makes user-based methods unscalable, since the number of users typically exceeds the number of items, which makes similarity calculations computationally expensive [178].

2.1.3. Hybrid Recommender Systems

RSs often face constraints when operating in isolation, particularly when diverse data sources are accessible [7]. Ideally, a RS uses all available knowledge from these sources and leverages the strengths of various recommendation algorithms to generate more robust and reliable recommendations [178]. To achieve this, *hybrid recommender systems* combine multiple paradigms [29]. Common hybridization strategies include (i) feature augmentation [211], where the output of one RS are used as input feature for the next, (ii) weighted techniques [255], where a score for an item is calculated using the weighted aggregates of the scores of the different RSs, and (iii) switching mechanisms [78], where the RS switches between different algorithms depending on the needs of a user. There are three main designs of hybrid recommenders: ensemble designs, monolithic designs, and mixed designs [7]. Ensemble designs combine results from existing algorithms into a single, more robust output, often using weighted averages or sequential processes where one algorithm's output informs another [7]. Monolithic designs

create unified recommendation algorithms by integrating diverse data types, sometimes modifying existing methods and blurring the lines between collaborative and content-based components. [7] Mixed systems present recommendations from multiple algorithms side by side, emphasizing the collective value of combined items rather than isolated suggestions [7].

2.2. Schwartz's Theory of Basic Human Values

The most widely accepted, validated, and tested model of values within psychology is *Schwartz's Theory of Basic Human Values* [185]. This theory defines values as universal motivational constructs and posits that a value's motivational goal is its defining feature. Values are seen as concepts or beliefs about desirable end states or behaviors that transcend specific situations, guide the evaluation of events and people, and are ordered by relative importance [186, 187]. The relative ordering of values by their importance to an individual or group is called a "value hierarchy".

This value hierarchy consists of an ordering of the ten universal value types[1]. These types are based on three fundamental human requirements: (i) biological needs, (ii) coordinated social interaction, and (iii) the welfare needs of groups. Everyone attaches importance to these values, but people and groups differ in how much importance they attach to each value, relative to other values [185]. Collectively, these values provide a framework for understanding human motivations across personal, social, and cultural contexts. Definitions for each value are given in Table 2.1

Table 2.1: Schwartz's Basic Human Values and Definitions [185]

Schwartz Value	Definition
Self-Direction	The pursuit of independent thought and action, emphasizing creativity, choice, and exploration.
Stimulation	The desire for novelty and excitement to maintain an optimal level of engagement.
Hedonism	The pursuit of personal pleasure and sensory gratification, prioritizing enjoyment and life satisfaction.
Achievement	The drive for personal success through demonstrating competence and meeting socially accepted standards.
Power	The pursuit of social influence, status, control over people or resources, and public recognition.
Security	Emphasizes safety, stability, and harmony within personal, relational, and societal domains.
Conformity	Involves restraining impulses and behavior to avoid harming others or violating social norms.
Tradition	Respect for cultural and religious customs, with motivation to accept and uphold inherited practices and beliefs.
Benevolence	Enhancing the welfare of close relationships through loyalty, helpfulness, and care.
Universalism	Commitment to understanding, appreciating, and protecting the welfare of all people and the natural environment.

These ten values fit together in a circular structure of relationships, according to their motivational compatibilities. This circular structure is called the circumplex. Values adjacent to each other express similar goals and are compatible (e.g., Power and Achievement both emphasize social superiority), while values on opposite sides of the circumplex represent conflicting motivations. An illustration of this circumplex can be seen in Figure 2.2.

Empirical testing of this theory found that the ten value types appeared as distinct regions in nearly all samples, with the circular structure of relations being replicated with high consistency across diverse cultural, linguistic, and religious groups [189]. Furthermore, testing found that these ten values are comprehensive, covering all the types of values that people attach importance to. It also indicated the existence of two higher-order bipolar dimensions. The first, Openness to Change versus Conservation,

[1] The theory has since been extended to consist of 19 distinct values. For this study, we continue to use the 10 values initially specified, as they have persisted through subsequent theoretical revisions

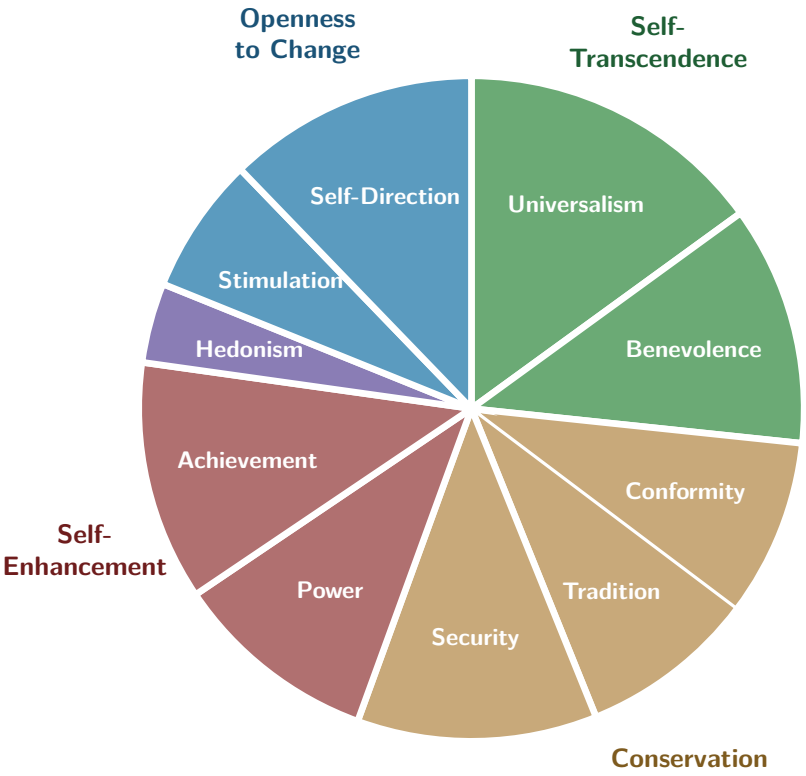


Figure 2.2: Circular structure of relationships for the 10 basic human values as adapted from Srivastava et al. [195]

contrasts the pursuit of novelty, creativity, and personal growth (Self-Direction and Stimulation) with the desire to maintain stability, predictability, and social order (Security, Conformity, and Tradition). The second dimension, Self-Transcendence versus Self-Enhancement, contrasts concern for the welfare of others and commitment to fairness and equality (Universalism and Benevolence) with the pursuit of personal success, power, and social dominance (Achievement and Power). These higher-order distinctions provide a conceptual framework for understanding the trade-offs and alignments among individual motivational priorities.

To measure values, Schwartz's Theory of Basic Human Values utilizes independent ratings of values [181], instead of directly asking individuals to rank the values by importance. Because individuals sometimes exhibit "scale use bias", a tendency to rate all values highly [217, 181], these ratings are standardized within an individual in a process called *ipsatization*. There are multiple techniques to implement ipsatization, but Schwartz's theory uses within-individual centering, in which an individual's mean rating is subtracted from their scores [185]. The ipsatized value scores are interdependent, so every value score includes information about all the other values in a system [181], which ensures that these ratings can be used to construct a value hierarchy. This creates a centered profile that accurately reflects an individual's unique motivational landscape.

3

Related Work

In this chapter we examine prior work that specifically addresses the integration of values in recommender systems (Section 3.1), as well as studies that explore value acquisition from natural language (Section 3.2). We only consider peer-reviewed articles written in English. These works either directly inform the design and methodology of the present exploration, or they represent approaches with limitations that motivate the contributions of this thesis.

3.1. Values in Recommender Systems

Historically, the majority of research into recommender systems has (implicitly) focused on system-centered values, which are values that apply to the system as a whole [18]. System-centered values reflect the strategic goals of the platform provider [113], the operational requirements of the algorithm, or the broader societal mission of an organization [19]. Given the fact that these systems are being constructed by designers who are used to thinking about systems, this focus is natural. When practitioners attempt to align these systems with user values, they frequently do so from this system perspective, measuring the alignment between system-centered objectives and user preferences [68, 67, 107, 253].

The fundamental issue with this approach is that measuring user preferences for system-centered values is not equivalent to measuring their actual values. Although they might have preferences about these system-centered values, the actual values that motivate their behavior are personal values [42]. Therefore, we need to investigate these personal values in order to get a complete understanding of value alignment in recommender systems. We summarize the existing work on this in Section 3.1.2.

3.1.1. System-centered Values

Examples of system-centered values are usefulness, fairness, and diversity [199]. These values are so prevalent in RSs research that they are sometimes called “standard” values [18]. Despite their prevalence, only a few works ([18, 199]) explicitly label these dimensions as values. Instead, research often refers to these values as “quality factors” [107], “performance metrics” [254], or simply “objectives” [10]. Notwithstanding the fact that they use different terms, all of these represent an entity that is a value.

Platforms believe that the primary reason users remain on a platform is the usefulness [199] of its recommendations [39]. Users are often in a relatively disadvantaged position with little control over algorithmic logic, so the platform’s objectives typically take precedence in the system’s design [128]. Therefore, the value that is embedded in the system design is usefulness. However, the value “usefulness” is too broad to be practical [19]. In order to optimize for this value, the designers of the RS specify what an abstract human value means in the context of a system and then translate that value into measurable objectives in a process called operationalization [56]. There are many ways to measure usefulness. The primary method through which the literature operationalizes usefulness is through accuracy metrics [89]. These metrics have become the standard evaluation metrics in RSs research [20], and have advanced research by providing a way to compare experiments. This has improved reproducibility [98], and aided the development of general-purpose recommendation algorithms [98]. Despite these impor-

tant achievements in accuracy metrics, it has long been recognized that accuracy on its own is often not a sufficient indicator of recommendation quality [147, 247]. This has led to the development of beyond-accuracy metrics, such as diversity [131] and novelty [33].

As a consequence of the realization that system-centered values might not reflect what users actually experience while interacting with the system [39], a new branch of RS research has developed that focuses on user-centric evaluation [170]. User-centric evaluations examine users' perceptions of recommender systems [84]. However, these evaluations often remain restricted to perceived versions of system-centered values, such as perceived fairness [192, 241]. The only difference with non user-centric research is that they investigate to what extent the user perceives that these values are exhibited by the system. However, they are still system-centered values in the sense that they are values about the workings and outcomes of the system.

Users differ in their preferences about the recommendations they would like to receive [67], and therefore about the system-centered values a system should exhibit. Informed by this fact, there has been research into aligning the amount of system-centered values a system exhibits with a user's preferences about those values. Commonly, these works investigate one value, for example, matching a user's novelty preference with the amount of novelty a system exhibits [112], or diversifying recommendations based on a user's diversity preference [59].

A small body of work has recognized that aligning the recommendations on individual system-centered values is not sufficient, since values relate to each other in the sense that increasing one value (i.e., usefulness) can decrease another value (i.e., diversity). To manage this trade-off, recent research employs multi-objective optimization (MOO) to simultaneously optimize for multiple values. These preference-based methods integrate individual user weights into the optimization process to ensure the final solution minimizes the distance to a user's unique preferences. Experimental results indicate that these multi-objective approaches can maintain stable accuracy while significantly improving diversity and novelty metrics [246, 76, 177, 237]. Furthermore, by effectively balancing these objectives, MOO can address challenges such as promoting long-tail items [233] and alleviating the cold-start problem for new users [234].

3.1.2. Personal Values

While system-centered values have been extensively studied, the role of personal values remains underexplored. Given that personal values are central to user behavior and preferences [130], this is a gap that needs to be addressed. Only a small body of research has investigated integrating personal values into recommender systems. These works use Schwartz's Theory of Basic Human Values (see Section 2.2) to operationalize values. Given that values are reflected in the language individuals use [27, 43], most of these works use free-text data sources to acquire a user's Schwartz values, such as tweets [117] and Amazon reviews [195, 96]. To acquire Schwartz values from these sources, these works use either a lexical approach [96], or the IBM Watson Personality Insights API [117, 195][1] The resulting extracted values are used for different purposes, such as preference classification and rating prediction [117, 96], identifying "gray sheep" users [195], and mitigating over-specialization [96]. While results show that incorporating value information can improve the effectiveness of RSs, they also indicate this improvement in effectiveness is only gained when value information is combined with user ratings.

Alternatively, a user can be asked to specify their value hierarchy themselves [23], mirroring the methods of self-report methods such as the Schwartz Value Survey. Research in the music domain indicates that incorporating user-specified value information in a RS increases engagement and improves self-reported satisfaction relative to a non-value-based baseline [83]. A similar strategy has been used by Jahanbakhsh et al. [97], who present an approach to value-aligning social media feeds. After specifying their value hierarchies, the authors use a LLM to rerank the items in social media feeds. Through two controlled studies, they demonstrate that users can effectively recognize and distinguish value-aligned feeds from standard engagement feeds. Value-ranked feeds diverged substantially from engagement feeds, with a near-zero rank correlation (mean Kendall's τ of 0.06), indicating that engagement metrics do not currently reflect what users actually want to see. Furthermore, while engagement

[1]The IBM Watson Personality Insights API has been discontinued in 2021 and is no longer available.

feeds favor Self-Enhancement values, value-aligned feeds significantly amplified content related to Self-Transcendence and Openness to Change.

3.2. Value Acquisition from Natural Language

Explicit value acquisition techniques such as surveys provide an accurate overview of a user's values, but they are time-consuming, intrusive, and prone to response biases [17]. As an unobtrusive way to acquire user values, researchers have investigated extracting personal values from users' written texts, as evidence indicates that values are embedded in the language that people use [27, 43]. By analyzing which words people use, these methods aim to gain insight into a user's values without explicitly asking them. The most common method for this is based on using a lexicon or dictionary that links specific words with certain values. These methods are called *psycholexical approaches* [42], due to their reliance on a lexicon. An early example of this approach is shown in Bardi et al. [17], where the authors built a small value lexicon based on the items from the Schwartz Value Survey (SVS), consisting of 3 words associated with each of the 10 values, which results in 30 words in total. Despite the small size, the extracted values based on this lexicon correspond with the results of the self-reported SVS, and with the behavior expected based on these values.

A much larger lexicon of 1068 words was constructed by Ponizovskiy et al. [169]. This "Personal Value Dictionary" combines words from three sources: the SVS, a value thesaurus by Christen et al. [42], and common unigrams. Based on five different single-authored, self-expressive text corpora[2], the authors refined and validated this dictionary. The dictionary is used to calculate scores for each of the 10 Schwartz values, the so-called PVD scores. To calculate these scores, they determine the rate at which words from each value type are used in a specific text. The correlation between these PVD scores and SVS scores is moderate for 7 out of 10 values, but the authors argue that small effects can still be important because they are visible in an intentionally inauspicious design.

[2]These five text corpora are: Corpus of Contemporary American English, Blog Authorship Corpus, CMU 2008 Political Blog Corpus, Essays on values and behaviors, and Facebook status updates.

4

Validation of the Personal Values Dictionary on the Goodreads Dataset

In this chapter, we validate if the method we use for acquiring user values can be reliably applied to the domain of our exploration, books. For value acquisition, we use the PVD by Ponizovskiy et al. [169]. This dictionary consists of 1080 English words, with their corresponding basic human value. This dictionary has been experimentally tested to correspond with self-reported values from the Schwartz Value Survey (SVS) in the domains of fiction, blog posts, essays, and Facebook status updates. However, in our exploration, we rely on book reviews. While reviews share characteristics with the types of text that the PVD has been tested on, they are not exactly the same. Therefore, we validate whether the PVD can also be applied to the new domain of (book) reviews, before continuing with our actual exploration.

For this validation we conduct experiments using two datasets containing book reviews: the Goodreads dataset from the University of California San Diego [226, 227], and the Amazon Reviews dataset [156]. We summarize the statistics of the two datasets in Table 4.1.

Table 4.1: Statistics of the Goodreads Poetry and Amazon Book Reviews datasets

Dataset	# Users	# Items	# Reviews
Goodreads Poetry Subset	47,400	36,514	154,555
Amazon Book Reviews	1,856,344	704,093	27,164,983

Goodreads is a book review site owned by Amazon, where users can review and store books. One of the ways to store the books is through the use of “shelves”, which are lists of books that a user has categorized in a certain way. The Goodreads dataset contains data scraped from `goodreads.com` users’ public shelves in late 2017. The authors of the original paper have released (1) metadata of the books, (2) user-book interactions (users’ public shelves), and (3) users’ detailed book reviews. Since the complete dataset is very large, the authors of the original paper provide medium-sized subsets, where the books are grouped by genre. We use the Poetry subset for this experiment, because this is the smallest subset available in terms of file size. This manageable size provides us with an opportunity to iterate through the experiments much faster, since we do not have to wait a long time for the experiment to finish.

The Amazon Review Data dataset contains reviews, product metadata, and links to products on `amazon.com`. The time frame of the review ranges from 1996 to 2018. The authors of the original paper provide smaller subsets for experimentation, which are organized per product category (e.g., “Fashion”, “Books”, “Electronics”). Since the domain of this experiment is book reviews, we use the Books 5-core

dataset, a subset of the Amazon Review dataset that includes data about book reviews that has been reduced such that all users have written at least 5 reviews, and each item has at least 5 reviews.

4.1. Metrics

To validate whether the PVD can be applied to the book domain, we apply the same validation metrics that the authors of the original paper used. An overview of these metrics can be seen in Table 4.2. We give a detailed explanation of these metrics below.

Table 4.2: Metrics for Evaluating the Personal Values Dictionary (PVD)

Metric	Definition
Temporal stability	Stability of scores over time, measured by correlating a user's value profile across time points.
Internal Consistency	Homogeneity within a test, ensuring items for the same construct are intercorrelated.
Convergent validity	Positive correlation between measures of conceptually related constructs.
Discriminant validity	Weak or negative correlation between measures of unrelated constructs.
Predictive validity	Ability of a measure to forecast future outcomes or behaviors.
Concurrent validity	Correlation between a new measure and an established standard for the same individuals.

These metrics are temporal stability, internal consistency, convergent validity, discriminant validity, predictive validity, and concurrent validity. Temporal stability describes the stability of test scores over occasions. This type of reliability is evaluated by correlating scores obtained for the same person at multiple points in time. Internal consistency is a description of the homogeneity within a test. Internal consistency is a description of the homogeneity within a test. For a hypothesized trait, such as dominance, the theory requires that items inquiring about behaviors subsumed under that label be generally intercorrelated.

Convergent validity and discriminant validity are different types of construct validity. Construct validity refers to the degree to which an operationalization actually measures the theoretical construct it is intended to measure [47]. Convergent validity is demonstrated when measures of a construct correlate positively with measures of conceptually related constructs. In the original paper, this was done by using Linguistic Inquiry and Word Count (LIWC) categories measuring constructs that are theoretically associated with specific values. Discriminant validity is a component of construct validity that measures whether the construct is easily discriminated from conceptually unrelated constructs. This can be demonstrated if the measure of this construct shows weaker or negative correlations with those conceptually unrelated constructs.

Predictive validity is a method for confirming that a psychological test or measure is useful because its results can successfully forecast a future outcome or behavior. Socio-demographic data of the users is required to test the predictive validity, which is not available in the dataset.

Concurrent validity tests whether a new test or measuring instrument is useful by comparing it against an established standard. Investigators administer a test, obtain an independent measure of the criterion on the same subjects, and then compute a correlation between the two. The Personal Values Dictionary was evaluated for concurrent validity by correlating value scores derived from text (PVD) with self-reported value scores obtained from the Schwartz Value Survey for the same individuals.

4.2. Experimental Setup

Due to the fact that we are conducting an offline evaluation, and no user-specific data is available in the dataset that would be required for convergent validity, discriminant validity, predictive validity, and concurrent validity, we focus on the temporal stability and construct validity metrics.

Similar to the original paper, we calculate the interrelations between the 10 values measured by the Personal Values Dictionary as a test of construct validity. The authors of that paper used multidimensional scaling to visually assess whether the spatial organization of the values replicated the value

circumplex specified by Schwartz. Multidimensional scaling is a statistical visualization technique that represents the similarities between objects as distances in a low-dimensional geometric space, typically two or three dimensions [129]. The goal is to produce a spatial configuration in which objects that are more similar to one another are placed closer together, and objects that are more dissimilar are placed further apart, thereby making the underlying structure of the data visually interpretable. In order to test the validity of the PVD on our dataset, we use multidimensional scaling on the four higher order values, similar to the original paper.

To calculate temporal stability, we construct a value profile for each user and compare the value profile of one year to the value profile of another year. These years do not have to be consecutive, since values are stable over time [93]. The value profile of a user is calculated by taking the mean of all the PVD scores for all the reviews from that year. We use these value profiles to calculate Pearson's correlation coefficient. For each value, a single correlation is calculated using all the users' value profiles. This statistic quantifies the linear association between users' values across the two years, ranging from 0 (no stability) to $+1$ (perfect stability across time).

4.3. Results & Discussion

The outcomes of the first experiment on the Goodreads dataset did not match the theoretical base. As can be seen in Table 4.3, the values of r for the values were between 0.0 and 0.1, which indicates very low stability. The mean of this experiment was 0.08, while the original paper reported a mean temporal stability of 0.32.

Table 4.3: Pearson correlation coefficients for temporal stability of different values on the Goodreads Poetry dataset

Value	**Pearson** r
Security	0.055529
Conformity	0.034052
Tradition	0.103561
Benevolence	0.119362
Universalism	0.093911
Self-direction	0.120609
Stimulation	0.056670
Hedonism	0.104838
Achievement	0.083107
Power	0.063748

The results of the multidimensional scaling we use for testing construct validity are visualized in Figure 4.1. In this figure, we plot the four higher order values (Openness to Change, Conservation, Self-Enhancement, Self-Transcendence) on a two-dimensional scale.

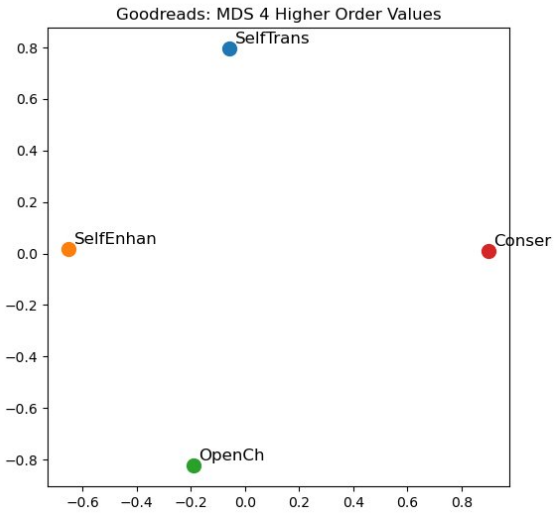


Figure 4.1: Outcomes of multidimensional scaling of the four higher order values in the Goodreads Poetry dataset.

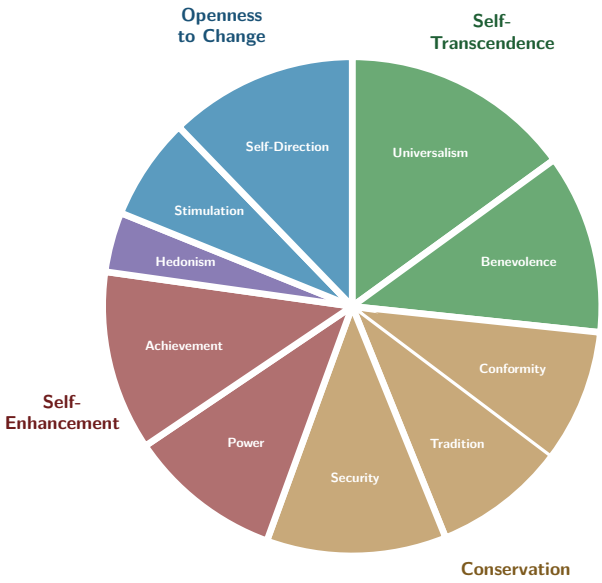


Figure 4.2: Circular structure of relationships among the 10 basic human values and the four higher-order values, adapted from Srivastava et al. [195].

The results of the multidimensional scaling show that the four higher order values form four different poles, which is consistent with Schwartz's circumplex. Since multidimensional scaling places objects that are more dissimilar further apart, these results indicate that the PVD is able to clearly differentiate between different values within the book domain. However, while the dissimilarity of the values matches the theoretical construct, the two bipolar dimensions we see do not. In the figure we see a Self-Transcendence and Openness to Change at opposite sites of the table, and a similar opposition for Self-Enhancement and Conservation. This does not align with the theoretically expected pattern visualized in Figure 4.2. According to that model, Self-Transcendence and Self-Enhancement should be on opposite sides, and similarly for Openness to Change and Conservation. This cannot be seen in our multidimensional scaling graph. We hypothesize that the lack of agreement with previous findings might be a result of our use of the Poetry subset. Poetry is a genre where values may be unstable due to the diverse and abstract nature of the writing. Therefore, we decided to use conduct results on an additional dataset to test whether our results were due to the genre.

For this we use the Amazon Review dataset, specifically the Books 5-core subset. This subset contains reviews of books for a wide variety of genres, instead of poetry exclusively. Using this dataset, the initial results were similar to the results of the Goodreads dataset shown above. Therefore, these results did not match the theoretical base as well. Upon reflection, we hypothesise that this might be the result of the experimental setup. The original experimental setup filters users based on their total number of reviews, with a minimum threshold of two. As explained in Section 4.1, value profiles are constructed by averaging the Personal Values Dictionary scores across all reviews written in a given year. A consequence of this setup was that users with only two reviews spread across different years (e.g., one review in 2003 and one in 2008) were included in the temporal stability calculation. A single review provides an insufficient basis for constructing a reliable value profile, since the PVD requires a minimum of 200 words to be effective. As a result, users with only one review per year could introduce considerable noise into the analysis and should probably not be considered meaningful data points for assessing temporal stability.

We test the effect of excluding these users using a method that filters out users who did not have a minimum number of reviews per year. We investigate different thresholds as the minimum number of reviews, ranging from one to seven reviews. We use seven as our maximum threshold, as manual inspection of the data indicates that the number of users who have more than seven reviews per year is virtually identical to the number of users who have at least seven reviews per year. The results of this can be seen in Figure 4.3.

Figure 4.3: Box plot of Personal Value Dictionary scores and amount of users per review count for the Amazon Book 5-core subset

As the figure illustrates, temporal stability increases substantially as the minimum review threshold rises, suggesting that users with more reviews per year exhibit more consistent value profiles over time. This indicates that applying a minimum review filter results in a more reliable estimate of temporal stability for this dataset. Applying this filter reduces the noise introduced by users whose sparse interaction histories may not provide a sufficient basis for constructing a stable value profile. Based on these findings, we test whether the filtering would have a similar effect on the Goodreads Poetry subset. The results of filtering on a minimum number of reviews per year can be seen in Figure 4.4

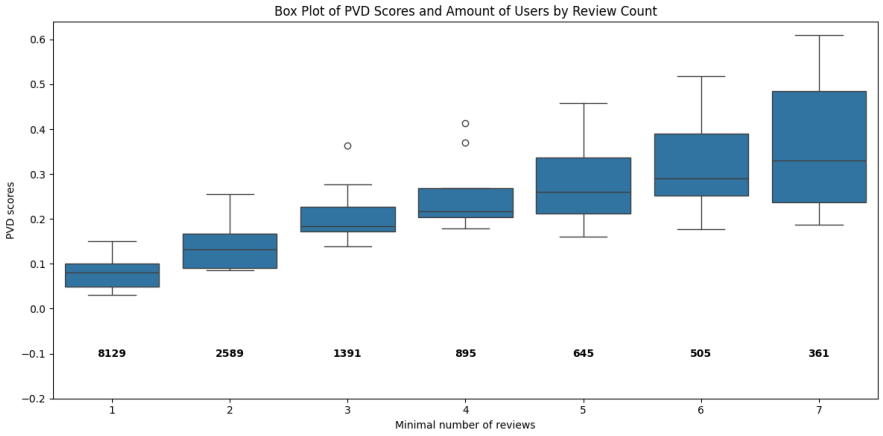


Figure 4.4: Box plot of Personal Value Dictionary scores and amount of users per review count for the Goodreads Poetry subset

As can be seen in the table, the same increase in temporal stability occurs for the Goodreads Poetry subset. The temporal stability reaches a mean that is comparable to the literature when filtering on users who have at least four reviews per year. Manual inspection of these results indicated that words from the title of the book were sometimes seen as value words. Since book titles frequently appear in reviews (e.g., “I loved The Power of Now”), words such as ”power” or ”freedom” in a title could artificially inflate certain value scores. Therefore, we test whether ignoring words from the title would result in a change in temporal stability. To achieve this, we preprocess each book’s title by lowercasing and removing all non-alphabetic characters, storing the resulting tokens in a lookup set keyed by book identifier. When computing the ipsatized value scores for a given review, we retrieve the title tokens for

the corresponding book and exclude any word that appears in both the PVD and the title token set. In this way, words that are associated with values are not counted toward the user's value profile when they appear solely as part of a book title mentioned in the review. The results of this can be seen in figure 4.5.

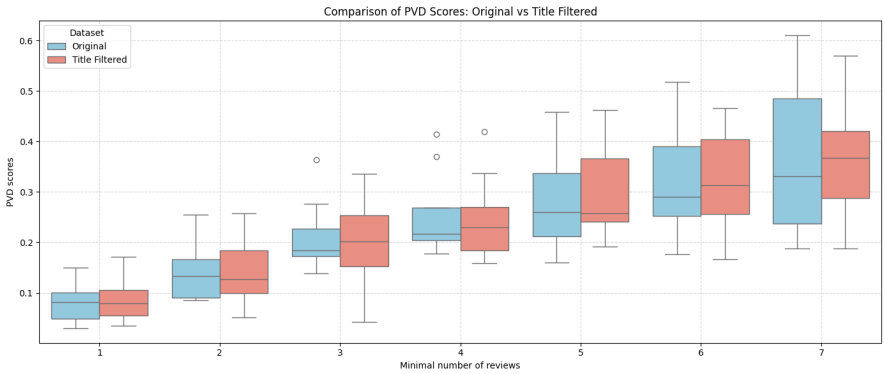


Figure 4.5: Box plot of Personal Value Dictionary scores and amount of users per review count for the Goodreads poetry reviews dataset, with a comparison between the case where the titles have been filtered, and where they have not been

These results indicate that filtering out title words did not meaningfully change the temporal stability of the value profiles. This suggests that, while title words can in principle inflate certain value scores, their occurrence is not frequent enough across the dataset to have a substantial effect on the aggregate temporal stability.

4.4. Conclusion

For construct validity, the clear separation of the four higher-order values into distinct poles demonstrates that the PVD is capable of meaningfully differentiating between value dimensions, which is consistent with the core structure of Schwartz's circumplex model. The results from the temporal stability analysis indicate that the temporal stability of users with at least four reviews is higher than for users with fewer than four reviews, with a mean that is comparable to the one found in the literature. This means that their value profiles are more stable. The way that this metric is calculated makes it logical that this is the case, since the temporal stability metric takes the average Personal Values Dictionary score for all the reviews of that year. Increasing the number of reviews likely results in a more accurate reflection of the value profile of the users, which in turn results in higher temporal stability values. Therefore, based on this exploration, we will filter users in our final experiment to have at least four reviews per year, to ensure valid and reliable metrics. Based on these results, we conclude that the PVD can be applied to the domain of book reviews.

5

Methodology

In this chapter, we outline the methodology we use to answer our research questions. Our research objective is to explore value alignment within recommender systems. To operationalize values, we use Schwartz's theory of Basic Human Values [185]. This theory identifies 10 basic human values that serve as motivational factors across all major cultures: Security, Conformity, Tradition, Benevolence, Universalism, Self-direction, Stimulation, Hedonism, Achievement, and Power. For a detailed description of the Theory of Basic Human Values, and definitions for each of the personal values, we refer the reader to Section 2.2.

Using this operationalization of values, we conduct an offline experiment using the Goodreads dataset [226, 227], which consists of data scraped from the book review site `goodreads.com` in late 2017. This public benchmark dataset [159] is commonly used in evaluating RSs [208, 145, 193]. Leveraging public benchmark datasets such as Goodreads enables us to compare our findings against existing studies in the field. This contributes to the development of a unified understanding of algorithmic performance and quality [89, 242], which is important for advancing knowledge of RSs. We focus on the book domain, because reading has been correlated with both personal values [15, 55] and overall life satisfaction [72]. We chose the Goodreads dataset over other book datasets (such as BookCrossing [252]) because it includes the (meta)data needed to enable our exploration. In particular, our exploration requires text written by a user and user-item interaction data to test whether a recommendation is relevant or not.

We start by generating recommendations from interaction data, without information about user values. According to our knowledge, no datasets exist with ground truth labels on personal values. Thus, we use the Personal Values Dictionary [169], a tool that maps 1080 English terms to their corresponding Schwartz values, as a way to add value information to the dataset. Using the PVD, we generate value profiles for both users and recommendations. These value profiles are used to assess the alignment between the values of the users and the values present in the recommendations. The resulting alignment acts as our baseline. Following this baseline analysis, we extend the recommender model by explicitly including user value profiles as features, produce a new set of recommendations, and re-evaluate the extent of value alignment. We then compare the baseline results with the results of the RS that used the value profile as features.

To ensure the reproducibility[1] of this work, we provide a detailed description of the different aspects of our methodology below. Following guidelines on reproducibility from prior work [21, 48], we report information about the dataset (Section 5.1), how we construct value profiles (Section 5.2 & Section 5.3), the algorithms we use (Section 5.4), the evaluation method (Section 5.5), and our implementation (Section 5.6).

[1] Reproducibility refers to the ability to independently verify the outcomes of a previous study by employing the original materials and methods [79, 242], which is essential for accountability in RS research [21]

5.1. Dataset

For our exploration, we leverage the Goodreads dataset from the University of California San Diego [226, 227]. Goodreads is a social book review site owned by Amazon, where users can review and store books. We summarize dataset statistics in Table 5.1.

Table 5.1: Statistics of the complete Goodreads dataset and the Goodreads English reviews subset, pre and post filtering

Dataset	# Users	# Items	# Reviews	Sparsity
Goodreads (multilingual)	876,145	2,360,655	15,739,967	99.98%
Goodreads English reviews	18,892	25,475	1,378,033	99.72%
Goodreads English reviews (filtered)	17,673	25,475	1,375,582	99.70%

The dataset contains data scraped from `goodreads.com` users' public shelves in late 2017, covering the period from January 2007 to November 2017 [213]. The core of the dataset consists of user-item interactions, which are recorded both as explicit and implicit feedback. The explicit data consists of ratings (ranging from 1 to 5) and users' detailed textual book reviews. The implicit data includes information about whether a book has been read, and whether a user has put this book on a "shelf", which is the site-specific terminology referring to a user-generated list. These shelves are used to organize, track, and manage a reader's book collection [205]. A shelf can encode implicit intent, since the most commonly used shelf names are "to-read" and "currently-reading" [144]. Lastly, the interactions are often associated with timestamps, recording when a user added, started, or finished a book.

The dataset also provides detailed (meta-)data on books. Next to basic identifiers such as titles, International Standard Book Numbers (ISBNs), and language codes, the dataset includes publication details such as publication year, release date, publisher, and the number of pages. The books in the dataset are associated with one or more genres.[2] These genres are derived from user-defined shelf names rather than a fixed taxonomy [226]. Each book is associated with multiple user-assigned genres and a count of how many users applied a particular genre label to that book.

To ensure methodological rigor and account for hardware constraints, we first conduct a preliminary experiment using the Poetry subset. This initial phase is designed to validate the applicability of the Personal Values Dictionary (PVD) to a new domain (book reviews) before conducting the full exploration. Following our findings in the preliminary experiment that the PVD is also effective with book reviews, we subsequently applied the method to the English review subset.

5.1.1. Data Filtering

We focus on the Goodreads English review subset, as the PVD was developed and validated only for English texts. This dataset was originally collected for spoiler detection [227], and is also referred to as the spoiler subset in the literature [194]. This is the only subset that consists exclusively of English reviews. All other subsets are multilingual, thus making them unsuitable for our exploration, as applying our extraction method to languages other than English would have unknown results.

The results of the preliminary experiment described in Chapter 4 indicate that the temporal stability of users' values reaches a comparable level to the literature when we filter on users who have at least 4 reviews per year. Therefore, we focus on this subset of users in our experiment. This aligns with data filtering techniques in the literature, where researchers commonly filter on users with at least 3 interactions to ensure the quality of the dataset [86, 111, 174, 152]. Statistics of the original and filtered Goodreads English review subset are shown in Table 5.1.

5.1.2. Data Imputation

The majority of the books in the dataset ($\sim 80\%$) have an associated description, which acts as a synopsis, summary, or "about" section for that book [144]. However, a large fraction of books is a missing description ($\sim 20\%$). Our method relies on book descriptions for constructing a value profile, so we investigated imputing missing descriptions by querying OpenLibrary to maximize the number

[2] These genres are: (1) Children, (2) Comics & Graphic, (3) Fantasy & Paranormal, (4) History & Biography, (5) Mystery, Thriller & Crime, (6) Poetry, (7) Romance and (8) Young Adult

of books with descriptions. OpenLibrary is one of the largest online library catalogs [165] and has been reported as an important book repository and a reputable website by prior work [166]. Using OpenLibrary only increased the amount of books with a description by 0.85%. Given this minimal increase in descriptions, we proceed without further imputation of missing descriptions.

A similar situation occurs with genre information. We use metrics that rely on genre information, and $\sim 40\%$ of the books in the dataset are missing genre information. To impute genre information, we use the Goodreads genre definitions by Thelwall [204] to extract genre information from the user-assigned genre tags in the dataset. This method halves the number of books missing genre information (from $\sim 40\%$ to roughly $\sim 20\%$), while improving the granularity of genre information, by using 37 instead of 8 genre labels. Therefore, we use these imputed genre labels in our exploration

5.1.3. Data Splitting

The choice of data splitting method can influence evaluation results [34]. We use a temporal split, as it is generally considered a good simulation of a RS's online behavior [148, 208]. Additionally, non-temporal splits cause data leakage [105], and are therefore not suitable. Values stay stable over time [210], meaning a temporal split does not negatively influence the validity of our experiment.

We use a 80% training set, 10% validation set, 10% test set. The validation set is necessary for hyperparameter tuning and to prevent overfitting [202]. This three-way split aligns with the splits used in related work [36, 139]. We treat the "date updated" of the review as the timestamp of that review [228], and select the most recent 10% of a user's interactions for the test set, as recommended by Jannach et al. [101]. We visualize this three-way split in Figure 5.1. Grouping by user ensures that the RSs are able to generate recommendations for all users, even when they are not active in the same timeframe. We illustrate our per-user splitting in Figure 5.2.

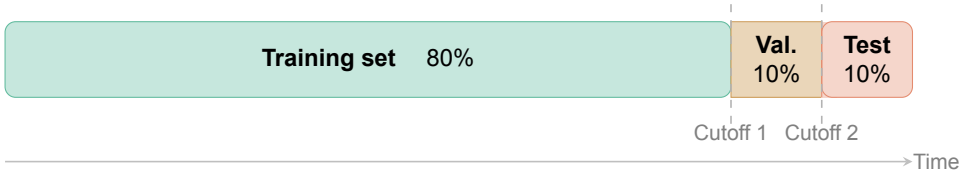


Figure 5.1: Temporal global split of the dataset into training (80%), validation (10%), and test (10%) sets. A single global cutoff timestamp separates training from held-out interactions; a second cutoff separates the validation and test sets. The "date updated" of each review is used as its timestamp.

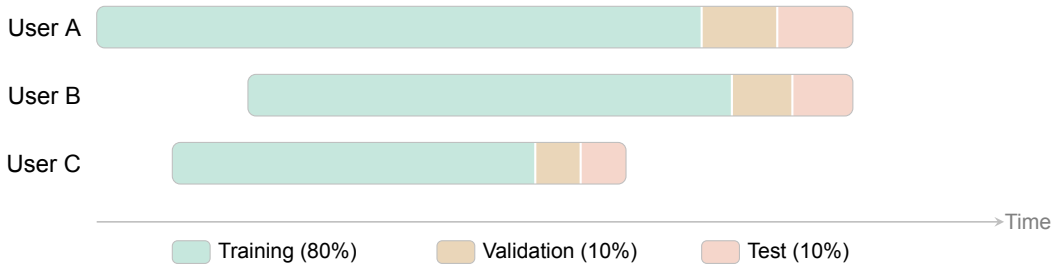


Figure 5.2: Per-user grouping of interactions. Each user's interactions are split such that the most recent 10% are assigned to the test set, the preceding 10% to the validation set, and the remaining 80% to the training set. Users may be active across different time windows; grouping by user ensures that recommendations can be generated for all users.

5.2. Constructing a Value Profile for a Text

To construct a value profile for a text we automatically assess mentions of personal values in a text using the Personal Values Dictionary (PVD) introduced by Ponizovskiy et al. [169]. This dictionary consists of 1080 English words, with their corresponding basic human value. Each word in the dictionary is associated with exactly one Schwartz value. This dictionary has been experimentally tested to correspond with self-reported values from the Schwartz Value Survey (SVS) [185], the validated self-report instrument used as standard in the literature [27]. These experimental results correspond with previous research that indicates a reference to values in a text can be seen as behavioral expressions of

corresponding values [17]. We use these extracted values as a ground truth about a user's personal values.

First, following common practice in text analysis, we preprocess the text by converting it to lowercase and removing non-alphabetic characters. We then load the Personal Value Dictionary and extract the value labels and word-to-value mappings. For each value category, we count word occurrences in the text using strict dictionary matching only. We calculate the final value score as the frequency of words representing the given value, minus the frequency of all value-related words in the text, formalized in Equation (5.1). We refer to this score as an ipsatized score, because it is a score that is the result of a process called ipsatization, where a score is standardized relative to an individual's own mean response across a set of items [181].

$$S_{ipsatized}(t, v) = f_{t,v} - \sum_{j=1}^{|V|} f_{t,j} \tag{5.1}$$

where $f_{t,v}$: frequency of words associated with Schwartz value v in text t; $\sum_{j=1}^{|V|} f_{t,j}$: total frequency of all value-related words in text t; V: the set of all Schwartz values ($|V| = 10$).

By performing ipsatization, we ensure that the resulting data reflect an individual's internal value hierarchy rather than just their general tendency to use value-laden language [169]. This is necessary, because individuals differ significantly in their general tendency to rate all items as highly (un)important regardless of their specific content, known as "scale use bias" [185]. This bias is also present in an individual's use of value-related words [169]. A positive score for a value v indicates that the user uses words associated with this value more frequently than their own overall value-language baseline, and a negative score indicates the reverse.

Combining the ipsatized scores for each value yields a value profile for a text, which consists of the 10 basic human values and their corresponding ipsatized scores. The value profile is represented as a vector, where each component aligns with the ipsatized score for one of the personal values considered in our exploration. Because the ipsatized scores are interdependent [181], each score indicates the relative emphasis of that value compared to other values in the text. Therefore, the value profile should be interpreted as a relative emphasis of the 10 human values in the text, compared to each other. The mean of each value profile is zero, as a consequence of the ipsatization process in which we center the scores to ensure that the sum of the scores is 0. A positive score for a given value (e.g., $+0.1$ for Self-Direction) indicates that this value is emphasized above the average level of value-laden language expressed in the same text, whereas negative scores indicate relative de-emphasis.

5.3. Constructing a Value Profile for Users & Recommendation Lists

To allow for easy comparison between the value profiles, we use the same method for constructing the User Value Profile (UVP) and the Recommendation List Value Profile (RLVP). We first aggregate all the relevant texts per user. The definition of "relevant text" varies depending on the objective. If the objective is to generate a user value profile, the relevant texts consist of all reviews authored by that user. For a recommendation list value profile, the relevant texts are of the descriptions of the recommended books. Subsequently, these relevant texts are concatenated into a single document. This approach mirrors the methodology used in the study that introduced the Personal Values Dictionary, wherein essay responses and Facebook data were pooled to form a unified corpus of behavioral samples [169]. Research indicates that profile correlations derived from short behavioral samples may exhibit low reliability [180]. By concatenating the texts into one comprehensive document, we treat the collection of texts as a single, substantial behavioral sample rather than as a series of short, isolated samples. This improves the reliability of the resulting profile. In this way, concatenation ensures that the rank ordering of values accurately reflects the user's true value hierarchy. To get the value profile for this document we apply the method described in Section 5.2. This results in the final User Value Profile/Recommendation List Value Profile.

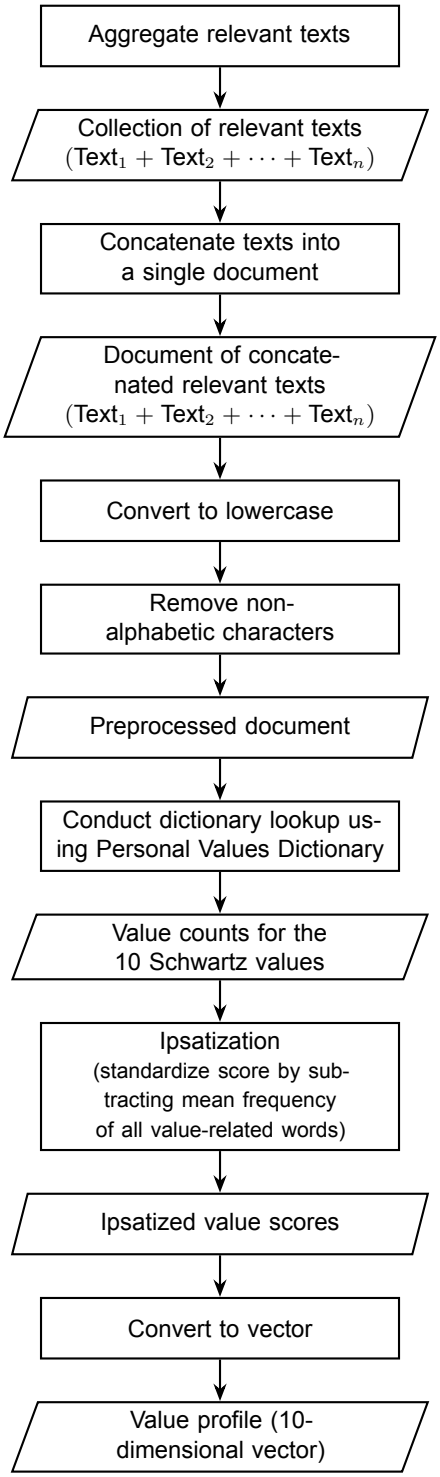


Figure 5.3: Pipeline for computing a value profile from text using the Personal Values Dictionary. Parallelograms denote data artefacts; rectangles denote processing steps.

5.4. Algorithms

The importance of considering baselines from different algorithm families has been repeatedly stressed in the literature [64, 63, 101]. Therefore, we use a set of algorithms that provides a representative sample of different RS approaches. We chose these algorithms because they are widely used [105] and provide competitive, strong baselines, which is important for assessing the performance of the

value-aware RSs [242]. Our main focus is on collaborative filtering (CF) algorithms, as this is the most popular and widely implemented technique in RS research [51]. In the following sections, we outline our reasoning for including each of these families, and how we insert the UVP into these algorithms. For background on each of these algorithm families, we refer the reader to Section 2.1.

We optimize these RSs through hyperparameter tuning, to avoid the issue of "phantom progress" [141, 242]. Phantom progress is a phenomenon where new recommender system models appear to outperform existing techniques, but these improvements are actually non-existent when the baseline methods are optimized [100, 242] Accordingly, we conduct a full hyperparameters grid search for each of these algorithms using all available hyperparameters provided by Recbole[3] on a dedicated validation split derived from the training set, with NDCG@10 as evaluation metric. We then use the best-performing hyperparameter configuration to retrain each model (both the regular and UVP variant) on the full training set before the final evaluation on the test set. This aligns with similar strategies used by prior work [152, 212]. We provide an overview of investigated hyperparameter values used in these searches in Appendix C.

5.4.1. Neighborhood-Based

Despite being one of the oldest methods, neighborhood-based methods are still one of the most popular methods for item recommendation [125]. This continued popularity can be explained by the fact that these methods are simple to implement, computationally efficient, and stable [157]. As representative of neighborhood-based methods we use User-based K-Nearest Neighbors (UserKNN) [175]. UserKNN is the standard user-based collaborative filtering method [86], which enables us to compare our results with prior work. We chose UserKNN over the item-based ItemKNN, as UserKNN allows us to utilize the user value profile to compute similarity scores.

UVP variant

To insert the UVP into UserKNN, we change the similarity function to also take into account value-based similarity between users. The resulting function (Equation (5.2)) is a linear combination of two similarity scores to determine the total proximity (Sim) between two users u and v. Similar work using this kind of combination in prior work on personality traits has shown promising results [158].

$$Sim(u, v) = \alpha \cdot \text{Sim}_{rating}(u, v) + (1 - \alpha) \cdot \text{Sim}_{values}(u, v) \tag{5.2}$$

where α: weight parameter that controls the contribution of value-based similarity in the overall similarity measure; $\text{Sim}_{rating}(u, v)$: rating similarity between users u and v; $\text{Sim}_{values}(u, v)$: value similarity between users u and v.

We use cosine similarity for both the rating similarity and value similarity scores, so that combining them is meaningful. Combining similarity scores derived from different metrics (e.g., cosine and dot product) would produce geometrically inconsistent results.

5.4.2. Latent Factor Models

Latent factor models have emerged as a dominant approach within CF, primarily because of their high accuracy and scalability [190, 235, 142, 239]. Because of these advantages, prior work has also investigated enriching these models with psychometric side information. These methods treat psychometric measures as additional user latent factors [207], and have been shown to improve recommendation accuracy [62]. Similar to personality, personal values are psychometric constructs, so inspired by this research we investigate the integration of personal values as latent factors in these models.

Among latent factor models, Matrix Factorization (MF) [124] stands out as one of the most widely recognized approaches. MF is a memory efficient method that excels at handling data sparsity [125]. The current state-of-the-art MF algorithm is Bayesian Personalized Ranking with Matrix Factorization (BPRMF) [173]. As the state-of-the-art for personalized ranking based on implicit feedback [14], BPRMF is one of the most cited papers in the field of recommender systems [87, 88, 213, 229, 152].

[3] Using the built in HyperTuning function in RecBole

BPRMF uses a pairwise learning-to-rank method, considering pairs of items to approximate the optimal ordering of a list, rather than predicting a single absolute score for an individual item [242]. Each pair is encoded as a binary preference label, where $+1$ indicates that the first item is preferred over the second, and -1 indicates the opposite. This transforms the problem into a binary classification task, in which the system aims to minimize the number of pairwise inversions in the training data. Because it focuses on the correct ordering of pairs, pairwise ranking loss is better suited for item recommendation than pointwise loss functions [152]. According to the 'missing not at random' hypothesis [196], user-item interactions are often missing because users have no interest in those items, and have consequently not interacted with them [140]. Pairwise ranking leverages this by treating unobserved items as "negatives" signals, allowing the model to learn effectively even when explicit negative signals are missing. This approach is valuable in sparse settings, where there are many unobserved interactions. Because the Goodreads English reviews subset is very sparse (99.7% sparsity), BPRMF is well-suited for our application. Therefore, we include BPRMF as representative of latent factor models.

UVP variant
To integrate information from the user value profile we use Equation (5.3), which is inspired by the equation used in personality-aware recommendation by Fernández-Tobías et al. [62].

$$\hat{r}_{ui} = q_i \cdot (p_u + W v_u) \tag{5.3}$$

where $\hat{r}_{ui}$: the predicted rating of user u for item i; q_i: item latent factor; p_u: user latent factor; v_u: vector representation of the user value profile; $W v_u$: linear projection of the continuous value vector into the latent space.

5.4.3. Context-aware

Context-aware recommender systems (CARS) are driven by the observation that user preferences are not static, but depend upon the specific circumstances or situation under which a recommendation is made [7]. By incorporating context as a third dimension alongside users and items, these systems achieve significant gains in recommendation accuracy [3]. Many different types of context have been considered in the literature, e.g., temporal context, social context, and spatial context. A context that is especially relevant to our exploration is psychometric context. Prior work has shown that including psychometric elements into CARSs can improve their performance [75]. For example, personality traits can be coded into context-aware architectures to enhance their predictive power [58]. Psychometric traits are used in CARS as a stable, long-term anchor that complements situational factors to refine user modeling and alleviate data limitations [184].

As a representative of CARS, we use DeepFM [81], a state-of-the-art representative of the deep click-through rate modeling paradigm [138] that combines deep neural networks (Deep) with Factorization Machines (FM) [172]. DeepFM is specifically designed to learn both low-order and high-order feature interactions simultaneously [81], leveraging deep neural networks to find non-linear patterns in user-item interactions [243]. DeepFM provides a benchmark for ranking-based tasks that refine the order of potential suggestions based on various contextual signals [245]. For these reasons, we chose DeepFM above other high performing approaches such as Wide & Deep (WDL) [40], which is limited by its reliance on expert feature engineering [81].

UVP variant
We integrate the value profiles as contextual factors in DeepFM. For this, we draw inspiration from previous work that integrated psychometric data as a "foundational user-related context" in CARSs [75]. Since DeepFM is already able to handle contextual features, no separate UVP variant is required.

5.4.4. Sequential

Many traditional recommendation algorithms are time-agnostic and ignore the chronological order of user interactions [101]. These methods assume that every past user interaction carries equal weight in determining their current preferences [230]. Breaking from the time-agnostic approach, sequential RSs assume that the temporal order of actions is often as important as the actions themselves [7]. For example, it makes sense to recommend coffee pods after a user buys a coffee machine, but not

the other way around. We include sequential models because these models are able to distinguish between long-term preferences and noisy trends in the data [7]. This is important for investigating personal values in recommenders, because these algorithms might capture these values as long-term preferences [36].

As a representative of sequential models, we use the Self-Attention based Sequential Recommendation model (SASRec) [111]. SASRec uses self-attention to adaptively find relevant items, leveraging the combined strengths of Markov chains [174] and recurrent neural networks (RNNs) [104]. Self-attention is a mechanism designed to capture internal relationships and dependencies within a sequence by matching representations against themselves [87, 164]. The self-attention mechanism of SASRec allows it to capture long-term and long-range dependencies within user interaction sequences, thereby overcoming the problems of RNN-based models (problems with long sequences), and CNN-based models (limited receptive fields) [220, 88]. SASRec has exhibited superior performance compared to earlier RNN-based (e.g., GRU4Rec) and CNN-based (e.g., Caser) models [25]. Due to its superior performance, it is a well-known representative of transformer-based sequential recommenders, which is frequently reported in the literature as a sequential baseline (e.g. [87, 164, 231, 232, 71, 88, 136, 139, 138, 143, 234]). For these reasons, we use SASRec as our representative algorithm for sequential recommendation.

UVP variant

Pure ID-based SRSs, such as SASRec, rely only on item ID and positional encodings [160]. While this has led to state-of-the-art performance, these SRSs suffer from data sparsity problems. A common way to mitigate this problem is to introduce side information into the SRS [41]. In our exploration, we integrate the user value profile as side information. The UVP can be seen as a form of long-term side-information, because personal values are stable over time [210].

We employ a late-fusion concatenation strategy [41] to integrate the static UVP with dynamic sequential signals, inspired by prior work on using psychometrics in SRSs [36]. By merging these representations at the prediction layer, we ensure a non-invasive [137] integration that prevents side information from distorting the self-attention mechanism [119]. This allows the model to simultaneously capture long-term stable preferences and short-term dynamic intents [36]

5.5. Evaluation

The main goal of our exploration is to measure the amount of value alignment of recommendations of RSs with user values. To evaluate this we measure the extent to which the value profile of recommended items matches the value profile of the user receiving them. Besides value alignment, we evaluate the RSs across a series of evaluation dimensions. Previous research has indicated that value-based RSs surface completely different items than standard (engagement based) RSs [97]. Motivated by these findings, we adopt a multi-metric evaluation framework in order to understand if and how incorporating value information into RSs changes the recommendations that are generated.

We use offline evaluation, which assesses a recommender system's performance using historical data rather than real-time user interaction [178]. The benefit of this method is that it provides high reproducibility, because researchers can use identical data, training/test splits, and statistical methods to verify and build upon prior work [101]. This enables systematic comparison of algorithms and settings without requiring real-time user interaction, allowing for controlled, quantitative analysis [242].

We treat the recommendation process as a ranking task. In ranking, the goal is to provide the user with a user-specific ranking for a set of items [173]. This differs from rating prediction, which optimizes for numerical accuracy in estimating user ratings [197]. Ranking is a more accurate proxy for user preferences than rating prediction, because minimizing rating error does not necessarily capture what users truly prefer or engage with [19], and because in real-life settings users are presented with ranked lists instead of predicted ratings [7]. Since it has been demonstrated that sampled evaluation produces inconsistent rankings [127], we use full-catalog evaluation, in which the RS is required to retrieve and rank the most relevant items for a user from the full catalog of items.

To evaluate the algorithms using rank-based metrics, we transform the explicit 5-point ratings into binary relevance labels by setting the rating threshold $\tau = 4$. This means that all items that are given a

rating of 4 or 5 are classified as items that a user finds relevant, while items with a rating of 1, 2, and 3 are considered non-relevant [242]. This choice is motivated by the fact that users use 5 star scales in an ordinal[4] instead of a cardinal[5] manner [126]. While a rating of 5 stars always indicates a higher relevance than 4 stars, the distance (and thus the difference in relevance) between the stars is not equivalent. In particular, 4 and 5-star ratings are closer to each other than 3-star ratings are to 4-star ratings [126]. This aligns with prior work that found that extreme ratings are more consistent across different trials than mild opinions (such as 3 star ratings) [12], which further motivates our choice to exclude ratings of 3 and lower. Besides these empirical reasons, using $\tau = 4$ allows for easy comparison with existing literature, as it is a frequently applied rating threshold when using data on a 1–5 star scale [6, 214, 215].

5.5.1. Metrics

To explore value alignment within RSs, we use metrics that capture both value alignment and the broad performance of the RS. For value alignment, we measure whether the value profile of recommended items aligns with the user value profiles. To characterize the broad performance of the recommender system, we include metrics that give us insights into different facets of the recommendation. This multi-metrics evaluation is seen as good practice for comprehensive evaluations of RSs [242]. Besides value alignment, we focus on the "key dimensions" [219] of the recommender system utility: accuracy, novelty and diversity. We complement these with metrics for coverage, fairness and popularity.

We include accuracy metrics, because they serve as the primary objective property for assessing a system's performance in the literature [214]. They are considered the *de facto* standard for comparing the performance of RSs, particularly in offline experimental settings [162]. This ubiquity is due to the fact that accuracy metrics allow researchers to compare various algorithms and configurations systematically at a relatively low cost, particularly in offline settings [242].

Furthermore, we report novelty and diversity metrics, because evidence suggest that users' Schwartz value profiles influence their preferences for novelty and diversity [96]. Specifically, prior research suggests that individuals with higher scores on self-direction, stimulation, and hedonism tend to prefer recommendations that are more novel [24]. Conversely, users who prioritize conservation or security may prefer more familiar and homogeneous items.

In order to understand the characteristics of the items that the RS is recommending, we include popularity, coverage, and fairness metrics. Popularity metrics give us insight into the popularity of a system's recommendations. Coverage metrics provide important insights into a system's reach and utility across the entire item catalog [74]. While accuracy metrics focus on how well a system predicts known interests, coverage evaluates the domain of items that the system is capable of exploring and recommending [89].

Coverage metrics are closely related to fairness metrics, and can be used to measure fairness over items, as they give us insight into the share of items or users that are served by the RS [242]. However, they do not give us insight into how (un)equally the items are recommended. To acquire that information, we use the Gini Index as our fairness over items metric.

Besides measuring fairness across items, it is also possible to compute fairness across users, such as metrics for group fairness. These metrics are used to investigate whether certain groups get consistently less relevant recommendations [242]. Because the Goodreads dataset lacks the data necessary to compute these metrics (e.g., demographic attributes [57]), we cannot compute these metrics. Therefore we do not consider them in our exploration.

Value alignment

The main focus of our exploration is the alignment between the ranking of values in the User Value Profile (UVP) and the Recommendation List Value Profile (RLVP). A value profile represents the relative ordering of the 10 basic human values. This relative ordering is a well-defined and comparable unit of analysis, as each profile consists of the same set of 10 values, all computed using the same ipsatized scale. This ipsatization standardizes the internal hierarchy, but removes information about the absolute

[4] Meaningful order, but the intervals between values are not consistent or measurable
[5] Meaningful order and consistent, quantifiable intervals

use of value words by the user. As a result, analyses that focus on single value dimensions (e.g. "Power") in isolation may be misleading. Accordingly, we compute similarity between users or between users and recommendation lists at the profile level, using a metric that compares the relative orderings rather than absolute values.

To evaluate the alignment between the ranking of values in the UVP and RLVP, we employ Kendall's τ_b rank correlation coefficient. This metric directly assesses the amount of agreement between the two rankings, which corresponds with our definition of value alignment.

Kendall's τ_b Kendall's τ_b rank correlation coefficient [114] measures the degree of agreement between two value orderings while explicitly accounting for ties.[6] This characteristic makes Kendall's τ_b particularly suitable for our analysis, as it effectively handles tied ranks in the data. This is important, because the value profiles frequently contain ties, as all the values that are not mentioned in the text will have the same score.

Values for τ_b range from perfect disagreement (-1) to perfect agreement ($+1$) between rankings. There is no universally accepted way to interpret the exact values of Kendall's tau [116]. In this evaluation, we use the following interpretation: values ranging from 0.00-0.10 indicate negligible agreement, 0.10–0.30 weak agreement, 0.30-0.50 moderate agreement, and values above 0.50 indicate strong agreement between two rankings [9]. This aligns with the recommended guidelines specified by Cohen [44]. We calculate the Kendall's τ_b between two value profiles using Equation (5.4), which is based on the formulation of Kendall [114].

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + T)(P + Q + U)}} \tag{5.4}$$

where P: #n concordant pairs; Q: #n discordant pairs; T: #n tied pairs in UVP; U: #n tied pairs in $RLVP$.

To calculate the agreement, Kendall's τ_b relies on the notion of concordant and discordant pairs. In our application of this metric to value profiles, a *concordant pair* is a pair in which the ordering of the values in the RLVP corresponds with the ordering of the values in the UVP, while a *discordant pair* is a pair in which the ordering of the RLVP does not correspond with that of the UVP. Kendall's τ_b compares all pairs of values in the two profiles, for example "Security" & "Conformity", "Security" & "Tradition", ..., "Achievement" & "Power".[7] In each comparison, it tracks how many pairs maintain the same relative order (concordant, P) versus reversed (discordant, Q). T indicates how many values have the same score in the UVP, but not in the RLVP. U indicates how many values have the same score in the Recommendation List Value Profile, but not in the User Value Profile. We visualize this in Figure 5.4.

[6] We chose Kendall's τ_b above other rank correlation coefficients, such as Spearman's ρ, because it can explicitly account for ties. While Spearman's ρ handles ties via rank averaging, it does not explicitly model them at the pairwise level, which Kendall's τ_b does.

[7] This results in $(\binom{10}{2} =) 45$ comparisons in total

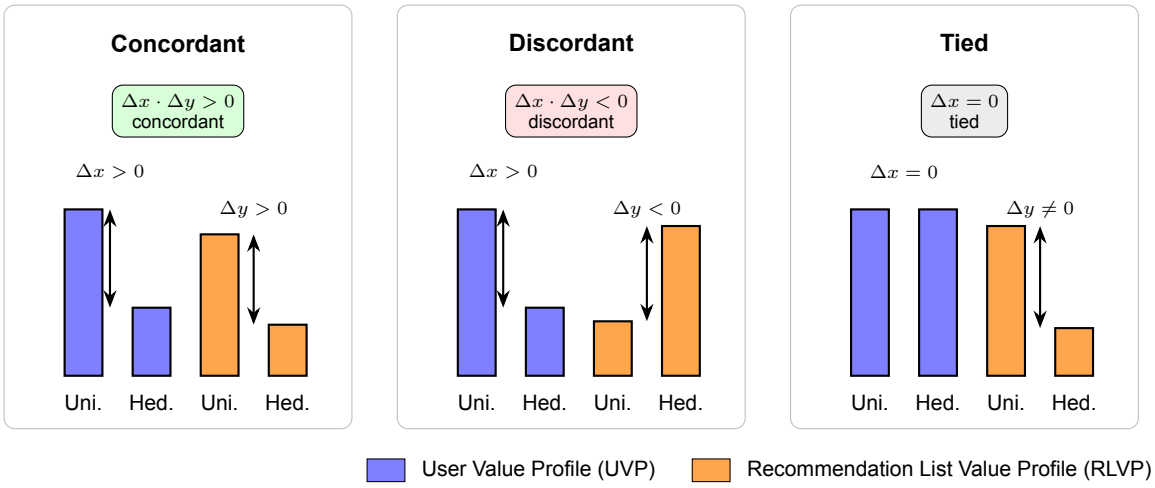


Figure 5.4: Illustration of concordant, discordant, and tied pairs in the context of Kendall's τ. Each panel shows a pair of values (Universalism and Hedonism) in the User Value Profile (UVP) and the Recommendation List Value Profile (RLVP). A pair is concordant when the relative ordering is preserved across profiles ($\Delta x \cdot \Delta y > 0$), discordant when it is reversed ($\Delta x \cdot \Delta y < 0$), and tied when one profile assigns equal scores to both values ($\Delta x = 0$ or $\Delta y = 0$).

Illustration Let $\mathbf{v}_u = [v_{u,1}, v_{u,2}, \ldots, v_{u,10}]$ be the UVP of user u and $\mathbf{v}_r = [v_{r,1}, v_{r,2}, \ldots, v_{r,10}]$ the RLVP of recommendation list r, each representing ipsatized scores over the 10 Schwartz values. We identify concordant and discordant pairs by comparing the relative ordering of pairs of values within each profile, rather than by comparing value scores at the same index across profiles. Equation (5.5) represents the ordering of a pair (i, j) in the UVP, and Equation (5.6) represents the ordering of the same pair in the RLVP.

$$\Delta x_{i,j} = v_{u,i} - v_{u,j} \tag{5.5}$$

$$\Delta y_{i,j} = v_{r,i} - v_{r,j} \tag{5.6}$$

A pair (i, j) is concordant if the relative ordering of the two values is the same in both profiles: $\Delta x_{i,j} \cdot \Delta y_{i,j} > 0$. To illustrate, assume that Universalism (denoted by $v_{u,i}$) is ranked higher than Hedonism (denoted by $v_{u,j}$) in the UVP. Since Universalism is ranked higher than Hedonism, $v_{u,i} > v_{u,j}$, and therefore $\Delta x_{i,j}$ will be positive.

If Universalism is also ranked higher than Hedonism in the RLVP, $v_{r,i}$ (Universalism) will similarly be higher than $v_{r,j}$ (Hedonism), so $\Delta y = v_{r,i} - v_{r,j}$ will be a positive number as well. Since Δx and Δy will both be positive, $\Delta x \times \Delta y > 0$, which means that a pair is concordant. This aligns with the definition of a concordant pair, since the relative ordering of Universalism and Hedonism is the same.

Alternatively, if Universalism is ranked lower than Hedonism in the RLVP, $v_{r,i}$ (Universalism) will be smaller than $v_{r,j}$ (Hedonism), so $\Delta y = v_{r,i} - v_{r,j}$ will be a negative number. Calculating $\Delta x \times \Delta y$ with Δx being positive, and Δy being negative results in a negative value, so $\Delta x \times \Delta y < 0$. This corresponds with the definition of a discordant pair, since the relative ordering of Universalism and Hedonism is reversed.

Finally, a pair is tied in either the UVP or RLVP if one of the value profiles ranks Universalism as equal to Hedonism, but the other value profile does not. This is the case if $\Delta x = 0$ or $\Delta y = 0$.

Accuracy

While investigating value alignment is our main goal, we also want to investigate the effect of value information on the accuracy of the RSs. To investigate how accuracy is affected by user values, we complement our value-alignment analysis with two standard performance-oriented ranking metrics:

Normalized Discounted Cumulative Gain [108] and Mean Reciprocal Rank [224]. These metrics complement each other: NDCG to assesses the amount of relevant items in the top of the ranking, and MRR assesses where the first relevant result is located.

Normalized Discounted Cumulative Gain Normalized Discounted Cumulative Gain (NDCG) is widely used in recommender system evaluation and is well-suited to settings in which recommendations are presented as ranked lists rather than as explicit rating predictions. Prior work [214] has shown that NDCG exhibits strong robustness and discriminative power when comparing ranking-based recommender models, making it an appropriate choice for assessing overall system performance. We use NDCG to capture the extent to which relevant items are ranked highly. It is important that relevant items are ranked highly, because users exhibit position bias [45]. Position bias is the disposition of users to interact with items on top of a list with a higher probability than with those in lower positions [106]. Therefore, these items should be relevant [108], and this is what NDCG captures.

The value of NDCG ranges from 0 to 1, where 1 indicates that the system ranks items exactly as they would be in an ideal scenario, with all relevant items at the top of the list according to their relevance scores. A value of 0 indicates that the system failed to recommend any relevant items, or that relevant items were placed so far down the list that their value was entirely lost due to positional discounting. We calculate NDCG using Equation (5.7).

$$\mathrm{NDCG} = \frac{\mathrm{DCG}}{\mathrm{IDCG}} \tag{5.7}$$

To calculate Discounted Cumulative Gain (DCG) we used Equation (5.8), where $2^{rel_i} - 1$ represents the gain value at position i, and $\log_2(i+1)$ implements the position-based discounting. The discounting reflects the assumption that the utility of a relevant item decreases with its rank position. Consequently, a model that fails to rank relevant items highly will receive a lower score. It should be noted that there are many different formulations for calculating DCG. We chose the formulation by Parapar and Radlinski [162], because this formulation uses a logarithm with base 2, which ensures all positions are discounted [178].[8]

$$\mathrm{DCG} = \sum_{i=1}^{N} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{5.8}$$

We calculate Ideal Discounted Cumulative Gain ($IDCG$) using Equation (5.9). The Ideal Discounted Cumulative Gain is the maximum possible DCG a system could achieve if all the relevant items were ranked at the very top of the recommendation list. We approximate this by taking all items known to be relevant to a user (i.e. items the user has interacted with) and sorting them in decreasing order of their relevance grades. We assume that all relevant items are placed at the top N positions, and apply the same logarithmic discount as in DCG.

$$\mathrm{IDCG} = \sum_{i=1}^{N} \frac{1}{\log_2(i+1)} \tag{5.9}$$

Mean Reciprocal Rank Mean Reciprocal Rank (MRR) [224] evaluates a recommender system's ability to place relevant items at the top of a recommendation list [215]. It is calculated as the average of Reciprocal Ranks (RR) across all users in the system [10]. For an individual user, the RR is the inverse (reciprocal) of the position (rank) of the first relevant item in their recommendation list [6]. The value for MRR ranges from 0 to 1, where a score of 1 represents a perfect system that always places a relevant item at the very first position [10]. We calculate MRR using Equation (5.10).

$$\mathrm{MRR} = \frac{1}{|U|} \sum_{u \in U} \frac{1}{\mathrm{rank}_u^*} \tag{5.10}$$

[8] The logarithm base can be chosen, since it is a free parameter. Typical values in the literature range between 2 and 10 [178]

where U: set of users; $rank_u^*$: rank position of the first relevant item found by an algorithm for a user u.

Diversity

Diversity refers to the internal variations within components of an experience [178], in particular to a collection of items and the degree of difference among them [110, 131, 219, 242]. It is important that RSs give the users a diverse list of items, because it gives them a wider array of options to choose from, thus increasing the chance that at least one of the recommended items will meet their needs [219]. Diversity can be measured from multiple perspectives. In our exploration, we focus on the user-level perspective. The user-level metric, Intra-List Diversity, captures the perspective of a user by measuring how similar or different items are within a single user's recommendation list

Intra-List Diversity We measure diversity using Intra-List Diversity (ILD), which is computed using the average pairwise distance of the items in the recommendations. ILD is a generally accepted metric in the literature, that has been shown to be a valid proxy for user-perceived diversity [102]. However, it should be noted that Jesse et al. [102] found that the specific implementation of the diversity metric is important for the diversity perception. Previous research indicates that users associate diversity with genre diversity [53, 218]. We therefore use the implementation that has been shown to exhibit the best match with user perception of diversity: genre-wise ILD [54]. The formula for this metric is given in Equation (5.11).

$$\text{ILD} = \frac{1}{|R|(|R|-1)} \sum_{i \in R} \sum_{j \in R} d(i,j) \tag{5.11}$$

where R: set of recommended items; i: first item of the comparison; j: second item of the comparison; $d(i,j)$: distance function.

As distance function, we use the genre information. Specifically, we use the complement of binary Jaccard similarity (Equation (5.13)) on genre information. This method is commonly applied in the literature [2, 102, 219], because genres can be treated as Boolean or binarized features, since an item either belongs to a genre or it does not [85]. This applies to book genres as well [144]. Hence, we use this similarity metric, instead of a similar metric such as cosine similarity, which is more suitable to numeric item features [33].

$$d(i,j) = 1 - J(i,j) \tag{5.12}$$

$$J(i,j) = \frac{G_i \cap G_j}{G_i \cup G_j} \tag{5.13}$$

where $J(i,j)$: Jaccard similarity between i and j; G_i: set of genres that i belongs to; G_j: set of genres that j belongs to.

Novelty

Novelty is an indispensable aspect of recommender systems, since the goal of a recommendation is discovery [219]. Discovery requires novelty, since recommending items that the user already knows defeats the purpose of the recommendation [2]. Following Castells et al. [33], we define novelty as difference between a user's present recommendations and their past experience. Therefore, we need a way to measure if a recommender system is able to recommend items that are novel to the user. For this, we use Unexpectedness, a user-dependent metric. We use a user-dependent metric, because these metrics align with our working definition that the novelty of an item depends on a user's previous experience [191, 102]. User-independent and popularity-based metrics such as global long-tail novelty do not align with that definition, and are therefore unsuitable for our exploration.

Unexpectedness We operationalize novelty using Unexpectedness, a metric that measures how surprising a recommended item is relative to a user's prior interactions [178, 191]. Unexpectedness reflects the difference between a user's present recommendations and past experience. We compute Unexpectedness using Equation (5.14), as formulated by Castells et al. [33]:

$$\text{Unexp} = \frac{1}{|R||I_u|} \sum_{i \in R} \sum_{j \in I_u} d(i,j) \tag{5.14}$$

where R: set of recommended items; I_u: set of items user u has interacted with; i: first item of the comparison; j: second item of the comparison; $d(i,j)$: distance function.

For each recommended item $i \in R$, the metric considers all items $j \in I_u$ that the user has previously consumed or interacted with. For every such pair (i,j) the distance function $d(i,j)$ measures how dissimilar the recommended item is from the user's past experience. Specifically, we compute $d(i,j)$ as the Jaccard similarity (see Equation (5.13)) between the genre information of items i and j. This formulation captures novelty as the extent to which recommended items belong to different genres than those previously encountered by the user, while remaining user-dependent through the incorporation of individual interaction histories. The choice to use genre information is supported by prior work that shows that utilizing genre information in novelty metrics is effective [168], specifically in the book domain [54].

Unexpectedness shares attributes with intra-list diversity (Section 5.5.1). Both use distance-based models to measure how different items are, but they compare those items to different sets [219, 178]. Unexpectedness compares the recommendations against the set of items the user has already interacted with, while ILD compares it against the set of recommendations itself. They thus capture different, but equally relevant aspects of the user experience.

Coverage

Coverage metrics quantify domain reach, measuring the extent to which the RS is capable of making recommendations over the available set of items [74]. This is important information, because a RS might appear successful by only recommending easy-to-predict popular items while ignoring the "long tail" of the catalog [89]. This in turn would prevent a user from discovering useful items they might not have found on their own, which impacts their satisfaction with the RS [191].

Item Coverage We use item coverage [89] as our coverage metric. Item coverage is sometimes also referred to as catalog coverage [99], since it measures how broadly the recommender system covers the item catalog in its recommendations. With item coverage, we measure the percentage of available items that are effectively ever recommended to at least one user over a specific period [74]. We calculate the item coverage using Equation (5.15). The values of this formula range from 0 to 1, where a result of 1 means the system is capable of surfacing every single item in the catalog to at least one person, while a score close to 0 indicates that the system's recommendations are concentrated on a small subset of the inventory [191].

$$\text{ItemCoverage} = \frac{\left|\bigcup_{u \in U} R(u)\right|}{|I|} \tag{5.15}$$

where U: set of users; R: set of recommended items; I: set of items.

Fairness

We include a fairness metric in our evaluation of recommender systems, because accuracy-based metrics fail to capture the social impact of a system [18]. In particular, we investigate fairness over items, which is also referred to as provider fairness [56]. Investigating fairness over items is important, because recommender systems operate in multi-stakeholder environments [199] where the interests of the providers of the items are just as relevant as those of the users [150]. While traditional metrics focus on user satisfaction, they often ignore whether the system's benefits are equitably distributed among providers of that content. To measure this, we include a fairness over items metric.

Gini Coefficient As a measure of fairness over items, we use the Gini coefficient [35]. The Gini coefficient is frequently used as a fairness metric because it measures distributional inequality [242]. We use the Gini coefficient because it is effective at detecting concentration biases [1, 99], the tendency of recommendation algorithms to focus their recommendations on a small part of the available item spectrum.

The Gini coefficient measures how evenly items are distributed among users in the recommendation list. It evaluates whether a recommender system over-recommends a subset of items or distributes recommendations more evenly across the catalog. The value ranges from 0 to 1, where 0 indicates perfect equality, in which all items are recommended equally often, and 1 indicates perfect inequality, in which a single item is recommended exclusively, and all others are ignored. We calculate the Gini Coefficient using Equation (5.16).

$$G = \frac{1}{n-1} \sum_{j=1}^{n} (2j - n - 1) p\left(i_j\right) \tag{5.16}$$

where n: total number of items; j:rank of the item; i_j: list of items ordered according to increasing $p(i)$.

Popularity

We include a popularity metric in our evaluation of recommender systems, because accuracy-based metrics alone do not capture the degree to which a system's recommendations are dominated by a small subset of highly popular items [18]. This phenomenon, commonly referred to as popularity bias [176], frequently occurs in CF methods [1]. Capturing this information is relevant for our exploration, as incorporating user value profiles as explicit features may change the distribution of recommended items in ways that are not reflected in accuracy metrics alone.

Average Popularity To measure popularity, we use the Average Popularity metric, defined in Equation (5.17), which computes the mean popularity of recommended items across all users.

$$\text{AveragePopularity} = \frac{1}{|U|} \sum_{u \in U} \frac{\sum_{i \in R} \phi(i)}{|R|} \tag{5.17}$$

where U: set of users; R: set of recommended items; $\phi(i)$: number of interaction of item i in training data.

5.6. Implementation

The computational environment for the experiments consists of a high-performance Acer Aspire 7 workstation. The CPU is a 12th Generation Intel Core i5-1240P, operating at a base frequency of 1.70 GHz, providing efficient multi-core processing for both general and specialized tasks. The system is equipped with 16.0 GB of RAM, offering sufficient memory for handling moderate-sized datasets and computational workloads. For accelerated computing, the workstation features a dedicated NVIDIA GeForce RTX 3050 Laptop GPU with 4 GB of VRAM, alongside integrated Intel Iris Xe Graphics (128 MB), enabling support for machine learning tasks.

We use the RecBole recommendation framework [249, 240, 248], because it allows for full experiment management, and has a clear configuration file that allows for experiments to be easily reproduced. We use RecBole's `significant_test.py` to run significance tests

For each recommender system, we select hyperparameters through systematic tuning. We report the resulting optimal configurations below.

UserKNN We conduct tuning over the neighborhood size k and the normalization parameter *shrink*. The optimal configuration was found to be $k = 10$ and *shrink* = 0.0.

BPRMF Tuning over the learning rate yielded an optimal value of $\lambda = 0.001$.

DeepFM The optimal configuration consisted of a dropout probability of 0.1, a learning rate of 0.01, and MLP hidden layer sizes of [512, 512, 512].

SASRec The optimal configuration consisted of an attention dropout probability of 0.2, a hidden state dropout probability of 0.2, a learning rate of 0.001, 2 attention heads, and 1 transformer layer. The item embedding dimensionality was set to 64, with an inner feed-forward dimensionality of 256. Weight initialization followed a normal distribution with a standard deviation of 0.02, and a layer normalization epsilon of 1×10^{-12} was applied for numerical stability. The `gelu` activation function was used in the feed-forward layer, and model training was conducted using the BPR loss function. All remaining parameters were kept at their default values.

6

Results

In this chapter, we report the empirical findings of our exploration, organized around the two central research questions. We first establish a baseline by examining the degree to which standard recommender systems already reflect users' personal values in their outputs (Section 6.1). We then report the effects of incorporating the User Value Profile across four recommender systems (UserKNN, BPRMF, DeepFM, and SASRec) evaluated along seven dimensions: value alignment, accuracy, diversity, novelty, coverage, fairness, and popularity (Section 6.2). In the remainder of this chapter, we discuss the patterns that emerge from these results in greater depth, including an analysis of popularity bias (Section 6.3), novelty (Section 6.4), fairness (Section 6.5), and the relationship between users' value profiles and the recommendations they receive (Section 6.8).

6.1. Baseline Value Alignment

Our first question aimed to gain insight into how user values are reflected in recommendation outcomes. Formally, we investigated RQ1: To what extent do recommendations generated by a standard recommender system align with users' personal values? In order to assess this alignment, we calculated the Kendall's τ_b correlation coefficient between the user value profile and the recommendation list value profile. The results of these calculations for each of the investigated models are shown in Table 6.1.

Table 6.1: Mean Kendall's τ_b correlation coefficients for different recommendation models.

Model	**Mean Kendall's τ_b**
UserKNN	0.2688
BPRMF	0.2690
DeepFM	0.1329
SASRec	0.2891
Overall Mean	0.2400

Both the overall scores and the isolated scores for these RSs fall between 0.0 and 0.3. Values for τ_b range from perfect disagreement (-1) to perfect agreement ($+1$), with values between 0.0 and 0.3 indicating weak alignment. Thus, the results indicate that recommendations of the standard RSs are weakly positively aligned with users' values.

6.2. Effect of Introducing the User Value Profile

Our next question investigated the effect of adding a user's values as input to the recommender systems. This was guided by RQ2: Does incorporating user values as features within a recommender increase this alignment? Besides measuring value alignment, we explore the broad performance of our chosen RSs using a multi-metric evaluation [242]. Our evaluation strategy includes metrics for

value alignment, accuracy, diversity, novelty, coverage, fairness, and popularity. These metrics cover all the main dimensions of evaluation, and taken together provide a comprehensive evaluation of the chosen RS. A detailed description of these metrics, our rationale for including them, and how they are calculated can be found in Section 5.5.1.

6.2.1. UserKNN

A comparison of the results for running UserKNN and the UserKNN + UVP variant can be seen in Table 6.2.

Table 6.2: Comparison of Evaluation Metrics: UserKNN vs. UserKNN + UVP (* indicates statistically significant difference, $p < 0.001$).

Dimension	Metric	UserKNN	UserKNN + UVP	% Change
Alignment	Mean Kendall's τ_b	0.2688	0.3267	+21.54%*
Accuracy	NDCG@10	0.0177	0.0096	-45.76%
Accuracy	NDCG@20	0.0231	0.0129	-44.16%
Accuracy	MRR@10	0.0244	0.0154	-36.89%
Diversity	Mean ILD	0.4741	0.5111	+7.80%
Novelty	Mean Unexpectedness	0.1384	0.2207	+59.46%
Coverage	Item Coverage@10	0.5803	0.3916	-32.52%
Fairness	Gini Index@10	0.8594	0.9427	+9.70%
Popularity	Average Popularity@10	349.9479	679.3777	+94.14%

The results, as shown in Table 6.2, indicate that the introduction of the UVP resulted in a significant increase in the mean Kendall's τ_b of 21.54%. We confirm this statistically via a paired t-test ($t(17599) = 30.16$, $p < .001$) and a Wilcoxon signed-rank test ($W = 56{,}749{,}268.5$, $p < .001$), both yielding a small effect size (Cohen's $d = 0.23$, 95% CI $[0.054, 0.062]$). We see an even greater absolute increase in mean unexpectedness, which increased by 59.46%. We also note small increases in intra-list diversity and the Gini index.

Besides a decrease in item coverage, we also note large decreases in the accuracy metrics MRR and NDCG. Because the NDCG metrics do not display a one-to-one correspondence with values previously reported on this dataset by Paparella et al. [161] (0.0984), we ran an additional experiment using the UserKNN implementation in the Elliot framework [13]. For that experiment, we used the same parameter settings as found in the code for the original paper.[1] A comparison between the results of the Recbole and Elliot implementations of UserKNN on the Goodreads English Review subset is shown in Table 6.3. We note that the NDCG scores from Elliot are similar to the metrics we found in running our experiment in Recbole. Therefore, we conclude that the UserKNN implementation in RecBole is not the cause of the lower NDCG.

Table 6.3: Comparison of Test Results for the UserKNN implementation of RecBole & Elliot on the Goodreads English Review Subset

Metric	UserKNN (RecBole)	UserKNN (Elliot)
MRR@10	0.0244	0.0603
MRR@20	0.0279	0.0659
NDCG@10	0.0177	0.0258
NDCG@20	0.0231	0.0309
Item Coverage@10	0.5803	0.5602
Item Coverage@20	0.7710	0.7300

6.2.2. BPRMF

We compare the results for the regular and UVP variant of BPRMF in Table 6.4.

[1] This code can be found here: `https://github.com/sisinflab/RecMOE`

Table 6.4: Comparison of Evaluation Metrics: BPRMF vs. BPRMF + UVP
(* indicates statistically significant difference, $p < 0.001$).

Dimension	Metric	BPRMF	BPRMF + UVP	Change (%)
Alignment	Mean Kendall's τ_b	0.2690	0.3524	+30.99%*
Accuracy	NDCG@10	0.0109	0.0104	-4.59%
Accuracy	NDCG@20	0.0141	0.0144	+2.13%
Accuracy	MRR@10	0.0178	0.0160	-10.11%
Diversity	Mean ILD	0.4911	0.4889	-0.45%
Novelty	Mean Unexpectedness	0.2195	0.2007	-8.57%
Coverage	Item Coverage@10	0.0600	0.2658	+343.00%
Fairness	Gini Index@10	0.9948	0.9647	-3.03%
Popularity	Average Popularity@10	1097.5635	751.8069	-31.50%

Introducing the UVP results in a statistically significant increase in mean Kendall's τ_b of 30.99%, confirmed by both a paired t-test ($t(17599) = 5.70$, $p < .001$) and a Wilcoxon signed-rank test ($W = 70{,}428{,}328.5$, $p < .001$), with a negligible effect size (Cohen's $d = 0.04$, 95% CI $[0.006, 0.013]$). Surprisingly, we see a decrease in accuracy metrics MRR@10 and NDCG@10, but an increase in NDCG@20. Another very large increase worth noting is the increase in Item Coverage. Besides these increases, we observe that unexpectedness and the Gini index decrease, and the intra-list diversity stays approximately the same.

Closer inspection of these results revealed that training times of these RSs differed greatly due to early stopping. BPRMF stopped after training for 20 epochs, while BPRMF + UVP ran for 70 epochs. To ensure that the observed difference in performance was not due to these differences in training time, we reran both RSs without early stopping. The results of this are seen in Table 6.5.

Table 6.5: Comparison of Evaluation Metrics: BPRMF vs. BPRMF + UVP
(* indicates statistically significant difference, $p < 0.001$).

Dimension	Metric	BPRMF	BPRMF + UVP	Change (%)
Alignment	Mean Kendall's τ_b	0.3359	0.3524	+4.92%*
Accuracy	NDCG@10	0.0105	0.0104	-0.95%
Accuracy	NDCG@20	0.0147	0.0144	-2.04%
Accuracy	MRR@10	0.0159	0.0160	+0.63%
Diversity	Mean ILD	0.4830	0.4889	+1.22%
Novelty	Mean Unexpectedness	0.1924	0.2007	+4.29%
Coverage	Item Coverage@10	0.3009	0.2658	-11.67%
Fairness	Gini Index@10	0.9565	0.9647	+0.86%
Popularity	Average Popularity@10	699.4952	751.8069	+7.48%

Comparing these results to those in Table 6.4, controlling for training time substantially changes the picture. The statistically significant increase in mean Kendall's τ_b persists, though the magnitude is considerably reduced from 30.99% to 4.92%, confirmed by both a paired t-test ($t(17599) = 12.57$, $p < .001$) and a Wilcoxon signed-rank test ($W = 65{,}195{,}001.5$, $p < .001$). This suggests that much of the alignment gain observed in the first comparison was an artifact of BPRMF + UVP receiving substantially more training. The dramatic differences in accuracy, novelty, coverage, and popularity metrics largely disappear when both models are trained for 70 epochs, with most metrics changing by less than 2%. Notably, the large increase in Item Coverage@10 reverses direction entirely, BPRMF now achieves higher coverage than BPRMF + UVP, further suggesting that the earlier coverage difference was driven by training time rather than the UVP component itself.

6.2.3. DeepFM

We compare the results of DeepFM and DeepFM + UVP in Table 6.6.

Table 6.6: Comparison of Evaluation Metrics: DeepFM vs. DeepFM + UVP
(* indicates statistically significant difference, $p < 0.001$)

Dimension	Metric	DeepFM	DeepFM + UVP	Change (%)
Alignment	Mean Kendall's τ_b	0.1329	0.1284	-3.38%*
Accuracy	NDCG@10	0.0036	0.0020	-44.44%
Accuracy	NDCG@20	0.0040	0.0032	-20.00%
Accuracy	MRR@10	0.0082	0.0041	-50.00%
Diversity	Mean ILD	0.6785	0.6525	-3.83%
Novelty	Mean Unexpectedness	0.4009	0.3805	-5.09%
Coverage	Item Coverage@10	0.0067	0.0060	-10.45%
Fairness	Gini Index@10	0.9992	0.9993	+0.01%
Popularity	Average Popularity@10	130.9624	98.1315	-24.99%

What stands out in the table is the fact that introducing the UVP into the recommender system decreased the values for all metrics except the Gini Index, which stayed the same. This includes a statistically significant decrease in mean Kendall's τ_b of 3.38%, confirmed by both a paired t-test ($t(17590) = -4.28$, $p < .001$) and a Wilcoxon signed-rank test ($W = 68{,}854{,}349.0$, $p < .001$), with a negligible effect size (Cohen's $d = -0.03$, 95% CI $[-0.007, -0.003]$). Another notable aspect is the very low accuracy scores in both the baseline and UVP variant. While surprising, prior work reports that in extremely sparse datasets the only pattern that DeepFM is able to learn might be popularity bias [121], which might explain these results.

6.2.4. SASRec

The resulting comparison between SASRec and SASRec + UVP can be seen in Table 6.7.

Table 6.7: Comparison of Evaluation Metrics: SASRec (136 epochs) vs. SASRec + UVP (127 epochs)
(* indicates statistically significant difference, $p < 0.001$).

Dimension	Metric	SASRec	SASRec + UVP	Change (%)
Alignment	Mean Kendall's τ_b	0.2891	0.3274	+13.26%*
Accuracy	NDCG@10	0.0250	0.0175	-29.96%
Accuracy	NDCG@20	0.0316	0.0235	-25.63%
Accuracy	MRR@10	0.0158	0.0121	-23.42%
Diversity	Mean ILD	0.5128	0.5043	-1.66%
Novelty	Mean Unexpectedness	0.3827	0.1876	-50.98%
Coverage	Item Coverage@10	0.8920	0.4799	-46.18%
Fairness	Gini Index@10	0.8570	0.9401	+9.70%
Popularity	Average Popularity@10	154.4449	246.3237	+59.49%

We observed a statistically significant increase in mean Kendall's τ_b of 13.26%, confirmed by both a paired t-test ($t(17599) = 16.70$, $p < .001$) and a Wilcoxon signed-rank test ($W = 65{,}183{,}127.0$, $p < .001$), with a small/negligible effect size (Cohen's $d = 0.13$, 95% CI $[0.034, 0.043]$). Besides the increase in Kendall's τ_b, we saw an increase in the Gini Index and a very large increase in average popularity. Next to a small decrease in intra-list diversity, we noted large decreases in the accuracy metrics NDCG and MRR, and even larger decreases in unexpectedness and item coverage.

6.2.5. Summary

In Table 6.8 we give an overview of the effect introducing the UVP has on each metric. This table shows that the significant gains in alignment are accompanied by systematic trade-offs in traditional RS metrics. In particular, ranking performance metrics tend to decrease when value information is introduced, while diversity- and novelty-related metrics show mixed shifts depending on the model. This pattern suggests that value-aware recommendation changes the optimization objectives.

Table 6.8: Comparison of Evaluation Metrics Across Models. Arrows pointing upwards indicate an increase, arrow pointing downwards indicate a decrease. Double arrows indicate a large increase or decrease (more than 20%).

Metric	UserKNN	BPRMF	DeepFM	SASRec
Mean Kendall's τ_b	↑↑	↑	↓	↑
NDCG@10	↓↓	↓	↓↓	⇓
NDCG@20	↓↓	↓	↓↓	⇓
MRR@10	↓↓	↑	↓↓	⇓
Mean ILD	↑	↑	↓	↓
Mean Unexpectedness	⇑	↑	↓	⇓
Item Coverage@10	↓↓	↓	↓	⇓
Gini Index@10	↑	↑	↑	↑
Average Popularity@10	⇑	↑	⇓	⇑

6.3. Popularity Analysis

Closer inspection of the data in Table 6.2 and Table 6.4 revealed a high popularity for UserKNN and BPRMF compared to the results obtained from DeepFM and SASRec. To explore the mechanisms behind this observation, we conducted further tests to identify the most frequently recommended items for each recommender system and their corresponding popularity.

We examined the most popular books in the dataset to gain deeper insight into whether the books that appear most frequently in the recommendation lists of the evaluated algorithms also correspond to generally popular items. We focused on understanding the relationship between algorithmic recommendation behavior and intrinsic item popularity. We defined a popular item as an item with a high number of positive interaction, where we used the same $\tau = 4$ as rating threshold for a positive interaction. The top 20 most popular books according to this definition can be seen in Table 6.9. We focused on the top 20 to focus on very popular books. Using this data, we analyzed the amount of overlap between the most frequently recommended books for each of the RSs and these most popular books.

Table 6.9: Top 20 Books by Positive Interaction Count

Book Title	Count
The Fault in Our Stars	2203
The Hunger Games	1909
Cinder	1736
Catching Fire	1353
The Martian	1336
Six of Crows	1298
Mockingjay	1276
Fangirl	1257
Scarlet	1255
A Court of Thorns and Roses	1231
Cress	1228
Eleanor & Park	1185
Daughter of Smoke & Bone	1129
Ready Player One	1104
Winter	1079
The Book Thief	1074
Insurgent	1055
Shadow and Bone	1054
Crown of Midnight	1029
A Court of Mist and Fury	1008

Table 6.10: Top recommended books for UserKNN and UserKNN + UVP. Highlighted titles are among the top-20 books by positive interaction count.

UserKNN		UserKNN + UVP	
Book	**Count**	**Book**	**Count**
Catching Fire	993	*The Fault in Our Stars*	3466
Mockingjay	964	*The Hunger Games*	2840
Scarlet	873	*Cinder*	2573
The Hunger Games	772	*Mockingjay*	2356
Cress	760	*Catching Fire*	1873
Shadow and Bone	711	*Insurgent*	1579
Queen of Shadows	702	*A Court of Thorns and Roses*	1557
Winter	676	*Scarlet*	1426
The Fault in Our Stars	670	*Eleanor & Park*	1416
Fairest	670	*City of Bones*	1354

Table 6.11: Top recommended books for BPRMF and BPRMF + UVP. Highlighted titles are among the top-20 books by positive interaction count.

BPRMF		BPRMF + UVP	
Book	**Count**	**Book**	**Count**
The Fault in Our Stars	8033	*The Fault in Our Stars*	4865
The Hunger Games	6806	*The Hunger Games*	3766
Cinder	6619	*Cinder*	3550
The Martian	5445	*Mockingjay*	3130
Divergent	5134	*The Martian*	2860
Ready Player One	5080	*Catching Fire*	2676
Catching Fire	5027	*The Book Thief*	1940
The Book Thief	4721	*Shadow and Bone*	1995
The Ocean at the End of the Lane	4311	*Fangirl*	1918
Insurgent	4305	*Insurgent*	1784

Table 6.12: Top recommended books for DeepFM and DeepFM + UVP. Highlighted titles are among the top-20 books by positive interaction count.

DeepFM		DeepFM + UVP	
Book	**Count**	**Book**	**Count**
Okay for Now	16018	*Words of Radiance*	16422
Beard Science	14843	*Okay for Now*	14158
Lumberjanes, Vol. 4: Out of Time	14699	*Beard Science*	13871
Long Way Down	13542	*The Hate U Give*	13576
The King of Attolia	12900	*Long Way Down*	13099
Words of Radiance	12741	*Roller Girl*	12040
Harry Potter and the Prisoner of Azkaban	12639	*The King of Attolia*	11997
Harry Potter and the Philosopher's Stone	11627	*Crooked Kingdom*	11787
The Way of Kings	9285	*Harry Potter and the Deathly Hallows*	7497
Allure	7500	*Nothing to Envy: Ordinary Lives in North Korea*	6824

Table 6.13: Top recommended books for SASRec and SASRec + UVP. Highlighted titles are among the top-20 books by positive interaction count.

SASRec		SASRec + UVP	
Book	**Count**	**Book**	**Count**
Harry Potter and the Cursed Child	1573	*Harry Potter and the Cursed Child*	3671
When Dimple Met Rishi	1337	*The Girl on the Train*	3370
Dark Matter	1231	*Six of Crows*	2756
The Upside of Unrequited	1144	*Crooked Kingdom*	2508
The Inexplicable Logic of My Life	884	*Uprooted*	2243
It Ends with Us	859	*A Court of Thorns and Roses*	2131
Our Dark Duet	850	*The Martian*	2073
The Girl on the Train	847	*It Ends with Us*	2061
Illuminae	735	*Caraval*	1880
A Crown of Wishes	684	*Wintersong*	1689

Table 6.10 and Table 6.11 show that UserKNN and BPRMF, in both baseline and UVP variants, exhibit substantial overlap between their most frequently recommended items and the top-20 most popular books. This finding indicates a strong popularity bias and is consistent with the literature, which suggests that neighborhood-based and matrix factorization methods often favor high-interaction items [99]. In contrast, DeepFM's recommendations in Table 6.12 shows virtually no overlap with the popular items, suggesting that its outputs are driven by learned feature interactions rather than raw item popularity. Interestingly, while these books are not very popular, DeepFM does recommend them to users extremely frequently. SASRec (Table 6.13) occupied an intermediate position: while its baseline did not show any popular items, the UVP variant recommends 3 more popular items. This observation may be taken to indicate that the UVP shifts recommendations toward globally popular content, which aligns with the rise in Average Popularity@10 seen in Table 6.7.

6.4. Novelty Analysis

Prior research suggests that personal values shape not only what users interact with, but how much novelty they prefer in their recommendations. Blomstervik and Olsen [24] found that individuals scoring higher on Self-Direction, Stimulation, and Hedonism tend to prefer more novel recommendations. To examine whether certain values are associated with higher unexpectedness, we computed a mean unexpectedness score per user by averaging across all items in that user's recommendation list. We then assessed the association between each Schwartz value dimension and this per-user mean unexpectedness using Spearman's rank correlation coefficient (ρ). Spearman's ρ was selected over Pearson's r for ordinal data because it measures monotonic relationships using ranks, avoiding the assumptions of linearity, normality, and equal interval spacing that Pearson's r requires but ordinal data violates. The results of this analysis are seen in Table 6.14.

Table 6.14: Spearman ρ between user value dimension scores and mean unexpectedness of recommendations, across all RSs. Cells marked n.s. indicate $p \geq 0.05$; all other values are significant at $p < 0.05$ or lower.

Value dimension	UserKNN	+UVP	BPRMF	+UVP	DeepFM	+UVP	SASRec	+UVP
Stimulation	0.144	0.197	0.206	0.200	0.059	0.040	0.045	0.187
Achievement	0.075	0.106	0.117	0.094	0.098	0.089	0.071	0.084
Universalism	0.105	0.084	0.084	0.098	n.s.	0.014	0.013	0.102
Self-Direction	0.056	0.078	0.070	0.102	n.s.	n.s.	n.s.	0.074
Conformity	0.055	0.072	0.077	0.054	0.093	0.064	0.048	0.061
Power	n.s.	n.s.	0.019	n.s.	0.064	0.041	0.019	n.s.
Tradition	-0.016	-0.019	n.s.	-0.017	-0.017	-0.029	-0.030	-0.019
Security	-0.029	-0.018	-0.025	-0.039	n.s.	n.s.	n.s.	-0.019
Hedonism	-0.085	-0.062	-0.082	-0.062	-0.050	-0.058	-0.050	-0.062
Benevolence	-0.293	-0.436	-0.400	-0.393	-0.280	-0.242	-0.189	-0.378

As can be seen from the table above, users who prioritized Stimulation, and Achievement received more unexpected recommendations, while those who valued Hedonism and Benevolence tended to receive less unexpected ones.

6.5. Fairness Analysis

The results of the fairness analysis showed that the overall mean Gini Index increased from 0.93 for the standard models, to 0.96 for the UVP variants. An overview of the findings is shown in Table 6.15.

Table 6.15: The results for Gini Index@10, grouped by model.

Model	Standard	With UVP	% Change
UserKNN	0.8594	0.9427	+9.70%
BPRMF	0.9948	0.9647	-3.03%
DeepFM	0.9992	0.9993	+0.01%
SASRec	0.8570	0.9401	+9.70%
Mean	**0.9276**	**0.9617**	**+3.67%**

Since a Gini Index closer to 1.00 indicates greater distributional inequality and concentration bias within a recommender system [215], this means that the item fairness decreases as the UVP is introduced. The reported values for the UVP variants are consistent with the results of Vaez Barenji et al. [213], who reported Item Exposure Gini scores between 0.95 and 1.00 for neighborhood-based and latent factor models.

6.6. Mean Value Scores in Recommendations

To get a better insight into the values expressed by the recommendation list value profiles, we calculate the arithmetic mean value scores for both the standard models, and the UVP variants. This gives us information about the average value profile of recommendations across the user population. Aggregating the ipsatized scores of multiple individuals to determine the overall relative emphasis for a group is a common and theoretically grounded practice in value research [42, 70]. Table 6.16 and Table 6.17 show the results for the standard and UVP models respectively.

Table 6.16: Arithmetic mean value scores for standard models

Value	BPR	DeepFM	SASRec	UserKNN	Overall
Benevolence	**0.0925**	**0.1228**	**0.1153**	**0.0921**	**0.1057**
Stimulation	0.0216	0.0434	0.0178	0.0216	0.0261
Self-Direction	0.0304	-0.0229	0.0330	0.0303	0.0177
Security	0.0044	0.0717	-0.0061	0.0049	0.0187
Power	0.0128	0.0271	-0.0009	0.0132	0.0130
Universalism	-0.0247	-0.0035	-0.0195	-0.0248	-0.0181
Achievement	-0.0192	-0.0351	-0.0185	-0.0192	-0.0230
Conformity	-0.0180	-0.0394	-0.0322	-0.0178	-0.0268
Hedonism	-0.0435	-0.0825	-0.0345	-0.0435	-0.0510
Tradition	-0.0564	-0.0817	-0.0543	-0.0567	-0.0623

Table 6.17: Arithmetic mean value scores for UVP variants

Value	BPR+ UVP	DeepFM + UVP	SASRec + UVP	UserKNN + UVP	Overall
Benevolence	**0.0941**	0.1023	**0.1019**	**0.0950**	**0.0983**
Stimulation	0.0247	0.0583	0.0231	0.0246	0.0327
Self-Direction	0.0411	-0.0284	0.0427	0.0346	0.0225
Security	-0.0110	**0.1045**	-0.0046	-0.0044	0.0211
Power	0.0089	-0.0151	0.0055	0.0122	0.0029
Universalism	-0.0159	-0.0191	-0.0178	-0.0224	-0.0188
Achievement	-0.0199	-0.0321	-0.0273	-0.0170	-0.0241
Conformity	-0.0235	-0.0429	-0.0273	-0.0201	-0.0284
Hedonism	-0.0362	-0.0616	-0.0393	-0.0413	-0.0446
Tradition	-0.0624	-0.0659	-0.0568	-0.0612	-0.0616

6.7. Genre-level Analysis

We examine what drives the difference in unexpectedness and intra-list diversity between the baseline and UVP variants of the selected models by conducting a genre-level analysis of the recommendations. For this, we calculate the novel genre ratio and the genre coverage. The novel genre ratio is the proportion of recommendation-list genres absent from the user's history. Genre coverage is the number of distinct genres in the recommendation list. The result of this analysis is shown in Table 6.18.

Table 6.18: Mean genre-level metrics per algorithm. Shaded rows indicate UVP variants

Algorithm	Unexpectedness	ILD	Novel genre ratio	Genre coverage
UserKNN	0.1384	0.4741	0.0444	18.31
UserKNN (UVP)	0.2207	0.5111	0.0762	19.73
BPRMF	0.2195	0.4911	0.0543	18.97
BPRMF (UVP)	0.2007	0.4889	0.0518	19.08
DeepFM	0.4009	0.6785	0.1748	22.89
DeepFM (UVP)	0.3805	0.6525	0.1705	21.56
SASRec	0.3827	0.5099	0.1295	18.77
SASRec (UVP)	0.1876	0.5043	0.0606	19.35

Table 6.19: Top 20 Books by Positive Interaction Count with Genres

Book Title	Count	Genres
The Fault in Our Stars	2203	chick-lit, contemporary, fiction, humor, romance, young-adult
The Hunger Games	1909	adventure, contemporary, fantasy, romance, science-fiction, suspense, thriller, young-adult
Cinder	1736	adventure, fantasy, magic, paranormal, romance, science-fiction, young-adult
Catching Fire	1353	adventure, contemporary, fantasy, romance, science-fiction, suspense, thriller, young-adult
The Martian	1336	adventure, contemporary, fantasy, humor, science-fiction, science
Six of Crows	1298	adventure, crime, fantasy, magic, mystery, paranormal, romance, thriller, young-adult
Mockingjay	1276	adventure, fantasy, romance, science-fiction, suspense, thriller, young-adult
Fangirl	1257	chick-lit, contemporary, fiction, humor, romance, young-adult
Scarlet	1255	fantasy, magic, romance, science-fiction, young-adult
A Court of Thorns and Roses	1231	adventure, fantasy, magic, paranormal, romance, young-adult
Cress	1228	adventure, fantasy, magic, paranormal, romance, science-fiction, young-adult
Eleanor & Park	1185	chick-lit, contemporary, fiction, historical-fiction, romance, young-adult
Daughter of Smoke & Bone	1129	adventure, contemporary, fantasy, magic, mystery, paranormal, romance, young-adult
Ready Player One	1104	adventure, contemporary, fantasy, humor, mystery, romance, science-fiction, thriller, young-adult
Winter	1079	adventure, fantasy, magic, paranormal, romance, science-fiction, young-adult
The Book Thief	1074	classics, contemporary, fiction, historical-fiction, history, literature, young-adult
Insurgent	1055	adventure, fantasy, romance, science-fiction, young-adult
Shadow and Bone	1054	adventure, fantasy, magic, paranormal, romance, science-fiction, young-adult
Crown of Midnight	1029	adventure, fantasy, magic, mystery, paranormal, romance, young-adult
A Court of Mist and Fury	1008	adventure, erotica, fantasy, magic, paranormal, romance, young-adult

6.8. Cluster Analysis of User Value Profiles

Personal values differ between individuals, and can be used to explain the difference between individuals [185]. An individual's value hierarchy differentiates them from other people. Presumably, a variety of different value hierarchies are present in the Goodreads dataset. Because value hierarchies are inherently individualistic, introducing value profiles into a recommender system need not benefit all users equally. This raised the question of whether alignment improvements are consistent across the user population, and whether users with different value profiles received systematically different recommendations. To investigate this, we perform K-means cluster analysis on the Schwartz value profiles of all users in the Goodreads English dataset. Following the methodology of Maslova et al. [146] and Lee et al. [134], we determine the optimal number of clusters using three criteria: the gap statistic, the silhouette coefficient, and the Calinski-Harabasz index (as a Python-equivalent of the Hubert index used by Lee et al. [134]). All three criteria converged on $k = 2$ as the optimal solution, yielding two clusters of comparable size ($n_1 = 9007$; $n_2 = 9885$). The value hierarchies of these clusters are shown in Table 6.20 and Figure 6.1.

Table 6.20: Mean Schwartz value scores per cluster (standardized prior to clustering; raw scores shown). Dimensions ordered by Schwartz circumplex position. Dominant values in **bold**.

Value Dimension	**Cluster 1** ($n = 9007$)	**Cluster 2** ($n = 9885$)
Self-Direction	0.090	**0.114**
Stimulation	−0.003	0.056
Hedonism	0.037	0.019
Achievement	−0.026	−0.008
Power	−0.042	−0.032
Security	−0.056	−0.059
Conformity	−0.054	−0.042
Tradition	−0.064	−0.065
Benevolence	**0.155**	0.039
Universalism	−0.037	−0.022

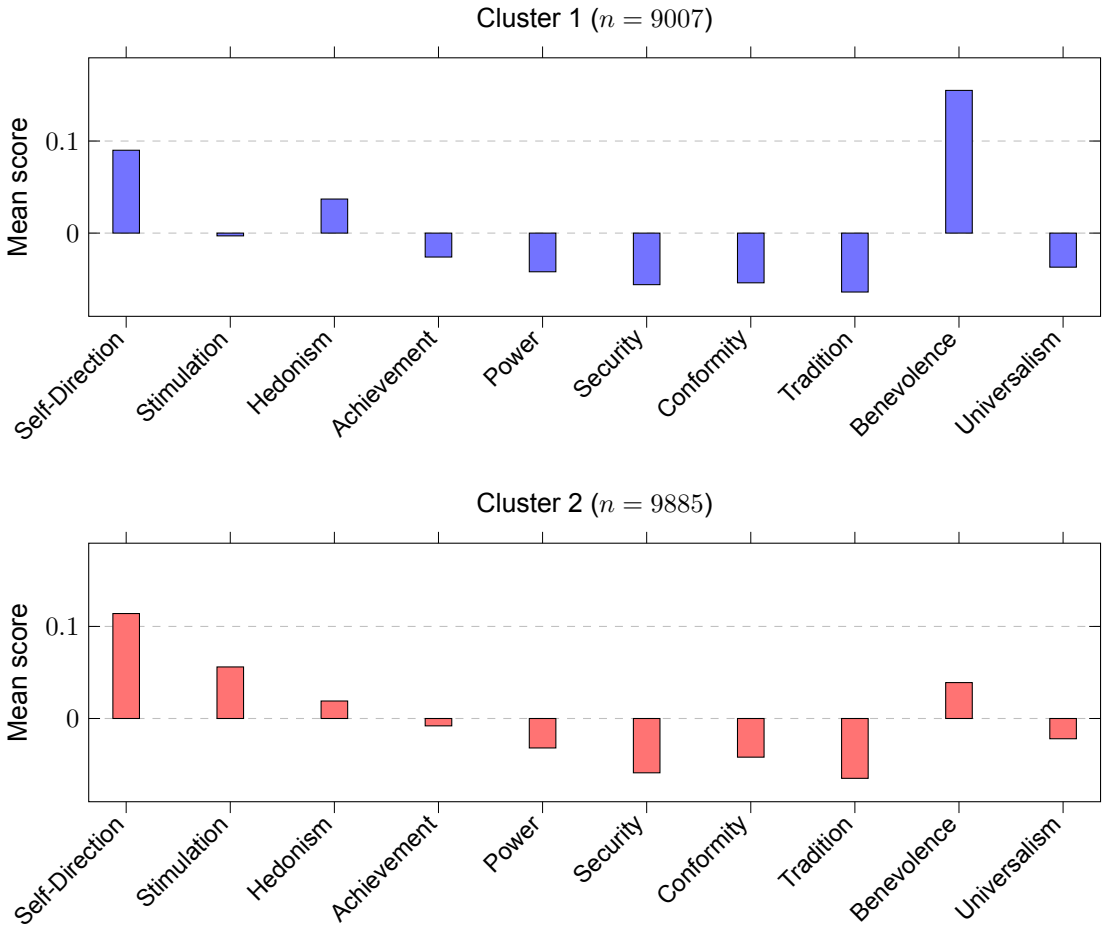


Figure 6.1: Mean Schwartz value profiles by cluster. Ipsatized scores are shown for each value dimension

We use these clusters to determine if the most dominant values in a user's value hierarchy influence the resulting alignment. We motivate this, by hypothesizing that, because the mean scores for Self-Direction are increasing when we insert the UVP (see Section 6.6), and the Self-Direction cluster is the biggest identified cluster, the alignment would increase, because all the scores are moved closer to Self-Direction. To determine whether there were any differences in value alignment per user, we investigated if there were any users for whom the alignment was always increasing, always declining, or a mix of those two. The results of this analysis are shown in Table 6.21.

Table 6.21: Cluster composition of consistency groups relative to the overall user distribution. The ratio column shows observed proportion divided by expected proportion under the null of no association

Group	Cluster	*n*	Obs.	Exp.	Ratio
Overall	1	8,483	48.2%	—	—
	2	9,108	51.8%	—	—
Always improve	1	629	46.4%	48.2%	0.96×
	2	727	53.6%	51.8%	1.04×
Always decline	1	387	47.0%	48.2%	0.97×
	2	437	53.0%	51.8%	1.02×

We found that neither the always improve nor always decline group deviates significantly from the overall distribution, indicating that value-cluster membership does not influence whether a user benefits from UVP.

7

Discussion

In this chapter, we discuss our exploration of how users' personal values are reflected in recommendation outcomes and whether the explicit incorporation of value information can improve alignment between these personal values and the values present in the recommendations. We compare our outcomes to the related work, highlighting consistent as well as contradicting findings. We also discuss societal and research-related implications of our findings.

7.1. Weak Alignment in Standard RSs

With respect to the first research question, we found that recommendations generated by standard RSs exhibit a weak degree of alignment with users' personal values. There are several factors that could contribute to this indirect alignment, most of which concern the relationship between Schwartz values and the interactions on which interaction data is based.

First, because values influence the actions of an individual [185], the explicit interactions captured by a RS can be seen as decisions driven by these personal values [82]. Based on this interaction data, the RSs might implicitly learn to optimize for these values, since they are associated with what a user interacts with. Second, prior research has identified strong correlations between specific value dimensions and content preferences [22, 117]. This means that when a standard RS aligns a recommendation with a user's past ratings, it could inadvertently align these recommendations with the values that motivated those ratings. Third, because values are acquired through socialization and shared experience, users tend to form behavioral clusters based on shared principles [82]. Standard CF might use these clusters to surface items that resonate with the collective values of similar user groups [195]. Finally, items sometimes contain linguistic and thematic features that correlate with specific values [117]. When a user engages with content that has these features, the system learns these features as preferences, though they reflect value-driven interests. In summary, while standard recommender systems do not explicitly model values, they may partially capture the behavioral effects of values embedded in the data. This could explain why their recommendations show a small positive alignment with the values that originally drove user interactions.

While standard RSs exhibit some positive alignment, the overall alignment is still weak. This lack of alignment indicates that the values of the recommendations do not reflect the user's values. We argue that this misalignment can have negative consequences for the users of RSs. RSs have a non-trivial influence on our behavior [77] through their central role on online platforms, where RSs structure the environment in which choices are presented to the user in many aspects of their daily life [151]. Through the recommendation process, RSs shape which options are visible and how they are prioritized [60]. Because there is no neutral way to present information [203], the resulting presentation of the recommendations nudges users toward certain behaviors [77, 103]. As a consequence, RSs shape what users see and how they experience the world [32], thereby embedding certain values [26]. If these embedded values diverge from the values of the individuals receiving them, the resulting behavior may similarly diverge from what individuals would choose for themselves [26, 31, 128]. As a result, their

behavior will not align with their values.

According to philosophical [179] and psychological accounts, such as Self-Determination Theory [167], living according to one's values is an essential requirement for well-being. Well-being is often considered the highest human goal [118], and promoting well-being is often seen as an ideal for technology [28]. Well-being is broadly understood as a condition in which individuals experience happiness and flourishing, such that their lives are going well for them [28]. If RSs nudge us into behavior that does not correspond with our values, this will negatively affect our well-being. Therefore, it is problematic that standard RSs do not align with our values.

Value alignment is also of critical importance for its own sake, as well as being necessary to ensure that RSs do not harm our well-being. It is essential that technology acts in line with our values, because that ensures that individuals can live according to the reasons and motives that they reflectively take as their own [50]. These reasons and motives were decided upon after a process of reflection where the individual recognized those values as a valid expression of their personal identity [50]. As such, personal values provide a much clearer image of what individuals really care about than the preferences that current RSs try to infer from interaction data [120]. Therefore, RSs should take these values into account.

7.2. Significant Increase in Value Alignment After Incorporating User Values

With respect to the second research question, the data revealed a significant increase in alignment when incorporating user values as features for three of the four RSs. This means that incorporating user values as features within a RS increases alignment of the recommendations with user values.

The only model that does not align with this general trend is DeepFM. Instead of increasing, the introduction of value information significantly decreases the value alignment between the recommendations and the users' values. There are several possible explanations for this result. First, DeepFM is known to underperform in extremely sparse datasets [121]. Since Goodreads is an extremely sparse dataset, this might be a possible explanation.

An alternative explanation for this result is that it is due to insufficient training times, which may have led to underfitting. Prior studies have noted the importance of sufficient training times for DeepFM [38]. Because the model has not yet learned personalized interactions it will recommend the same small subset of items to almost every user. This explanation is supported by the extremely low item coverage we see in the results, and by manual inspection of the results, which show that DeepFM recommends a small amount of books much more frequently than the other RSs.

Finally, integrating static value information might not be suitable for use in CARSs. While static psychometric information has been integrated into CARSs previously, prior work has stressed that contextual information should be a characteristic of that context, instead of a static user characteristic [3]. An important characteristic of personal values is that they are context-independent. While this makes them applicable to every context, it also means information about these values is not a characteristic of that specific context. Thus, it is possible that personal values are used as static user characteristics in DeepFM. This might explain the declining performance in the UVP variant of DeepFM. Based on these explanations we conclude that the decrease in value alignment for DeepFM, while significant, is not indicative of the general effect that introducing user values has on RSs.

The significant change in value alignment, coupled with the weak value alignment for standard RSs, shows that while standard RSs implicitly incorporate information about values, these values are not modeled as an explicit optimization target. This might result from the fact that standard RSs are evaluated and optimized according to metrics that have no relationship with personal values. This is supported by prior work that found that value-based rankings are uncorrelated with engagement based rankings [97].

7.3. Trade-Offs between Value Alignment and Other Metrics

Coupled with increasing value alignment, we found that incorporating value information into RSs affects the other metrics in a variety of ways. This means that while incorporating information about values significantly increases value alignment, it also comes with corresponding trade-offs across the other evaluation dimensions. Below we discuss the findings per dimension, since the exact effect that introducing value information has differs.

We found a consistent trade-off between value alignment and accuracy. When alignment improved, accuracy metrics decreased across all models. A possible explanation is that the inclusion of value-based signals introduces a separate optimization objective that cannot be fully aligned with the interaction-based ranking goals within typical model frameworks. This finding is contrary to previous studies which have suggested that using information about personal values can improve the accuracy of the recommendations [117, 96]. This discrepancy can be partially be explained by the fact that previous work determined accuracy using error-prediction metrics, while our work used ranking metrics. These metrics are based on different underlying assumptions [215], and high performance in error-based metrics does not guarantee similar performance in ranking-oriented evaluation [46]. Thus, an improvement in error-prediction metrics is compatible with a decline in ranking-based metrics [6]. If we shift the comparison to focus on ranking-based metrics only, we find that our results are consistent with the literature. Using the ranking-based metric Mean Average Precision, prior work found that combining psychometric information with rating information decreased accuracy [96]. This aligns with the decrease we found.

The results for diversity and novelty are mixed: UserKNN and BPRMF show improvements on both dimensions following the introduction of the UVP, while DeepFM and SASRec show declines. With respect to diversity, the decreases observed in DeepFM and SASRec are consistent with prior work finding that adding personal value information increases intra-list similarity and thus reduces intra-list diversity [96]. The improvements in UserKNN and BPRMF suggest that simpler CF approaches may respond differently to value information. In the case of UserKNN in particular, the UVP may have surfaced items from adjacent value clusters within the user neighborhood, thereby increasing variety rather than narrowing it. Notably, the overall change in intra-list diversity across all models was small, which suggests that value integration does not dramatically reshape the diversity of recommendation lists.

We see a similar pattern for novelty, with UserKNN and BPRMF showing increases in unexpectedness and DeepFM and SASRec showing decreases. For the latter two models, the introduction of value profiles appears to pull recommendations closer to a user's established preferences, thereby reducing novelty. The clearest example of this is the large decrease in unexpectedness for SASRec. At baseline, this model already achieved high unexpectedness, presumably by capturing long-range sequential patterns in user interactions. The introduction of the UVP may have introduced a persistent signal that overrode the model's exploratory tendencies. This tension is consistent with prior work: sequential models like SASRec are designed to capture dynamic preferences and long-range dependencies [238], while value profiles represent stable, long-term anchors [206]. The stable, context-independent nature of the UVP may therefore act as a static constraint on the dynamic recommendation process. This explanation is supported by prior findings of Xie et al. [238], who noted that the method of integrating psychometric information in sequential recommenders has a significant effect on the resulting system's performance.

Interestingly, we found that for UserKNN using the value profile as an additional input feature resulted in a large increase in unexpectedness. This increase might be a consequence of implementation of integrating the UVP into UserKNN. We integrated the UVP into UserKNN by replacing the rating-based similarity with a linear combination of rating similarity and value profile cosine similarity [236]. This means that two users can now be considered neighbors not only because they have rated the same items similarly, but also because their value profiles are similar. Importantly, users with similar value profiles need not have overlapping consumption histories [16]: a user who strongly values Stimulation may share that orientation with users whose reading histories are quite different from their own [250]. When recommendations are generated from such neighborhoods, items from outside the user's immediate consumption cluster can enter the list, increasing the distance between recommended items and the user's own history. The increase in unexpectedness is therefore probably a direct consequence of the similarity function modification.

The introduction of user value profiles has consistent effects on coverage, fairness, and popularity, suggesting that value-aware recommendations come at a cost to distributional diversity. With respect to coverage, we found that incorporating the UVP generally reduces the range of items recommended across users. This indicates a degree of overspecialization in which the recommendation focus narrows to a smaller subset of the item catalog. This narrowing is accompanied by a reduction in fairness over items: both our fairness metric and manual inspection of the recommendation lists reveal greater distributional inequality and a stronger concentration bias following the introduction of value information. This aligns with prior research that also found an decrease in fairness after introducing psychometric information [96]. Combined with this pattern, we found that average popularity of recommendations increases after introducing the UVP. Taken together with the increase in the Gini coefficient this suggest an increase in popularity bias. A possible explanation for this consistent pattern across all three metrics is that when an algorithm is tasked with aligning recommendations with a specific user value profile, it may converge on a limited set of high-interaction items that represent safe matches for that profile [89]. Popular items might be more likely to contain sufficient value-relevant signal to satisfy the alignment objective, causing the system to favor them at the expense of less prominent items.

Overall, these results indicate that the impact of value integration depends upon the model's underlying mechanism, and that this effect is not uniform across evaluation dimensions. We observe consistent trends across all four RSs for accuracy, item coverage, fairness, and popularity, suggesting that incorporating the UVP influences these dimensions regardless of the underlying architecture. By contrast, the results for diversity and novelty are mixed, with no consistent direction of effect across models. Since these two dimensions are related, both conceptually and in the way we have implemented them in our exploration, these results suggest that the relationship between value integration and item discovery is more architecture-dependent.

Besides indicating that values and traditional metrics exhibit a trade-off, we argue that these results show that the values that RSs are traditionally optimized on, are conceptually completely distinct from personal values. RSs are evaluated and trained according to a wide variety of traditional metrics. These metrics operationalize concepts such as usefulness, fairness, and diversity [199]. Research often refers to these concepts as "quality factors" [107], "performance metrics" [254] or simply "objectives" [10]. Notwithstanding the fact that they use different terms, we argue that all of these concepts represent values. This perspective is also taken by recent work investigating the role of values in RSs (e.g., [18, 199]). We contend that these values should be seen as the values that RS practitioners have identified as relevant values in the design of RSs. When constructing and designing RSs, practitioners decide which values to prioritize and how to operationalize these values into technical requirements [39]. Although correlations between personal values and the values prioritized by RS researchers, designers, and practitioners exist [24], the two remain conceptually distinct. Values that are important in designing RSs do not need to have any relationship with the personal values that motivate who a person is and what they do. This distinction carries practical consequences. When recommender systems are optimized for such metrics, there is a real risk that they are optimizing for the wrong things. In optimizing for traditional metrics, these RS are aligning with "top-down" defined notions of value rather than the values that actually matter to the people they are designed to serve.

In order to ensure that RSs are aligned with personal values, these systems should explicitly incorporate personal values as an optimization target [120], with corresponding operationalizations. The current metrics used in RS research are operationalizations of a completely different value system. As long as personal values are not an optimization target, evaluating and training RSs with traditional RS metrics will not bring us closer to aligning the recommendations with user values.

In our exploration, we have used Kendall's τ_b as evaluation metric to measure value alignment between the value hierarchy of the user, and the value hierarchy of the recommendations. By doing so, we have shown one possible operationalization of value alignment that future researchers could use as an optimization target. We do not claim this is the only or optimal metric for measuring value alignment. Our research only shows that it is possible to use such a metric, and implement it in popular RS models. Fortunately, the psychological literature on personal values provides many options to measure values. These metrics have been extensively empirically validated. Herlocker et al. [89] identified key considerations that RS practitioners should keep in mind when selecting metrics for their evaluation. These include the comparability of results with other published research, the validity of underlying assumptions,

sensitivity to detect real differences, and the threshold for statistical significance. A substantial portion of metrics currently used in RSs research fails to meet these requirements [46, 242]. For personal values, RSs researchers can draw from the extensive insights in psychology to meet these requirements. The validity of underlying assumptions [185], sensitivity to detect real differences [182, 188], and threshold for statistical significance [66, 153] have all been extensively studied within psychology. Therefore, future researchers would not have to conduct additional research to answer these questions, but can use these insights from psychology. Additionally, this helps RSs researchers ground and contextualize the results more broadly than traditionally possible, since they can compare their work to both published research in RSs, and published research in psychology.

7.4. Schwartz Value Clusters in the Goodreads Dataset

Through cluster analysis, we identified two clusters in the Goodreads dataset, broadly characterized by Benevolence-oriented and Self-Direction-oriented value profiles. This clustering differs from previously reported four- and three-cluster solutions in general population studies [146, 134]. This divergence can be explained by the nature of our dataset. Goodreads users constitute a self-selected population of heavy readers [205], and prior research indicates that such populations exhibit different value hierarchies than the general population, with stronger emphasis on self-direction, universalism, and benevolence [55, 109]. This corresponds with the clusters that we have identified. Therefore, we suggest that the identified clusters are a consequence of this self-selection process.

Interestingly, we found that even in the standard RSs, recommendation behavior differs across clusters. Self-Direction-oriented users tend to receive more diverse and unexpected recommendations, whereas Benevolence-oriented users exhibit higher baseline alignment. A possible explanation of these findings is that since CF uses the relationship between users to make recommendations, standard CF might use these clusters to surface items that resonate with the collective values of similar user groups [195]. These results further indicate that values are already partially encoded in interaction patterns, which is consistent with our finding of weak positive value alignment.

Concerning value alignment, we found no clear difference between clusters, with neither cluster having a significantly higher proportion of users for whom alignment increases or decreases. This suggests that membership of a specific value cluster is not the reason for the individual-level gains in alignment. Instead, the improvements are more likely due to the RS responding to individual differences in user profiles. This is corroborated by closer inspection of the data. This inspection revealed that users with high scores on the Power value were more likely to experience a decline in value alignment, which indicates systematic misalignment when value profiles are introduced. The reason for this might be that Power is more difficult to infer reliably from behavioral or textual signals [169]. In our experiment, noisy or weakly grounded value estimates may shift the recommendations further away from the user's actual values. In contrast, users who score higher on Stimulation, Universalism, and Hedonism tend to consistently benefit from value integration. It seems probable that these results stem from the fact that these value dimensions are better captured by the PVD and thus generate more reliable value signals.

Closer inspection of individual values showed that the values Stimulation, Achievement, Universalism, Self-Direction, and Conformity are associated with higher novelty. Users who have value hierarchies that promote these values tend to receive more novel recommendations. This finding is mostly consistent with that of Blomstervik and Olsen [24], who found that individuals scoring higher on Self-Direction, Stimulation, and Hedonism tend to prefer more novel recommendations. For Self-Direction and Stimulation, we find the same pattern. However, for Hedonism, we find the opposite: the value of Hedonism is associated with less novelty. A possible explanation for this might be that the users who value Hedonism in this dataset also value values that are associated with low unexpectedness.

7.5. Research Implications

Prior work has outlined three necessary stages to align intelligent systems, such as RSs, with human values [120]. These three stages are (i) value acquisition, (ii) turning values into an optimization target, and (iii) training a model to optimize for this target. In our exploration, we have implemented each of these stages, thereby illustrating one possible way in which value alignment can be achieved. Future researchers can use these insights in the design of value aligned RSs.

For value acquisition, we have shown that psycholexical analysis can be successfully applied to the free-text book reviews. This extends the literature by indicating that these reviews can be seen as self-authored texts that contain information about personal values. Through our implementation, we have shown that the PVD can be successfully applied to the new domain of book reviews.

Based on the acquired values, we have empirically demonstrated that user values can be incorporated into RSs at an early stage of the recommendation process. Previous research into incorporating information about personal values has mostly relied upon reranking [96, 97] and matching methods [82, 83]. Since reranking methods only apply to the recommendations generated by a RS earlier in the pipeline, these methods are limited in the extent in which they can surface certain values. If the recommendations generated by the pipeline do not contain any items with values that a user attaches importance to, these methods are unable to value-align the recommendations. As a consequence, the effect of incorporating values at an early stage of the recommendation process was unclear. Our findings show that incorporating this information at an early stage is possible, and can lead to significant improvements in alignment. This is important for future researchers, as it shows that values can be used as an optimization target within RSs.

This exploration further extends the existing literature by demonstrating four distinct approaches to achieving early value integration, each tailored to a different family of recommendation algorithms. It is important to note that these approaches represent only one possible realization of early integration. We do not claim that these methods constitute the optimal or only means by which such integration can be achieved. Nevertheless, the fact that early integration is both feasible and yields encouraging results carries meaningful implications for future research on incorporating values into recommender systems.

7.6. Societal Impact

While our findings are limited to the book domain, we believe that they have many important implications for society and the role of RSs in that society. Both the weak amount of value alignment in standard RSs, as well as the increase in value alignment after incorporating value information are instructive in this regard. Furthermore, the relationship between value alignment and the metrics we have used also give us important insight into both the way values influence recommendations, and about the metrics themselves.

We argue that the identified disconnect between personal values and the values that traditional metrics are based upon also applies to other ways that have been proposed to include human values in RSs. All of these efforts are very valuable, and could improve the way that RSs and individuals interact. However, these methods still apply a collection of values that might not align with the values of an individual.

The fundamental problem with these approaches is that they take a “one-size fits all” approach to integrating values into RSs. A small part of existing research recognizes that users value different things, instead of being a homogeneous group with identical values. This research tries to match the values a system exhibits with a user’s preferences about those values. In their most basic form, these works investigate one value, for example matching a user’s novelty preference with the amount of novelty a system exhibits [112], or diversifying recommendations based on a user’s diversity preference [59]. Other works investigate matching multiple values to user preferences simultaneously, since values relate to each other in the sense that increasing one value (i.e., usefulness) can decrease another value (i.e., diversity). These methods use multi-objective optimization to simultaneously optimize for multiple values [246, 76, 177, 237]. This is usually achieved by integrating individual user weights into the optimization process, to ensure the final solution minimizes the distance to a user’s unique preferences [233].

However, this research still suffers from the problem that these values have been defined from a “top-down” perspective: RSs practitioners have identified that diversity is an important value that RSs should exhibit, and they correspondingly try to infer the diversity preference of users. While increasing diversity might be relevant for RSs practitioners, and for the multi-stakeholder systems in which RSs interact, it is no guarantee that a specific user values fairness.

Moving beyond the book domain, there are other domains in which RSs are applied, which might be more closely connected to the core of who an individual is. An important domain in which RSs are used is online dating. Values directly and indirectly influence partner preferences [80, 201]. For example, individuals are drawn to partners with similar levels of Universalism, Tradition, Hedonism, and Conformity [135]. In such cases a lack of value alignment might be more problematic than a lack of value alignment in the book domain. Our findings show that it is possible to significantly increase value alignment by incorporating information about values, thereby providing a possible way to improve value alignment in domains that are more closely connected to the core of an individual.

7.7. Ethical Statement

Using personal values in RSs raises ethical concerns that researchers must carefully consider. Since personal values are the core of one's personal identity [90], data about values is among the most sensitive psychometric information that can be collected. This is compounded by the fact that value profiles can be inferred from behavioral traces, meaning users can be profiled without their awareness or consent. Beyond privacy, values-based profiling can be used for manipulation: as demonstrated by incidents such as the Cambridge Analytica scandal, psychometric data can be exploited at scale to influence behavior and public opinion [200]. These concerns do not argue against the use of personal values in recommender systems, but they do underscore the need for transparency, and careful governance of how value information is acquired, stored, and applied.

7.8. Limitations

While this study provides meaningful insights into value alignment in recommender systems, we acknowledge several methodological limitations. These limitations relate primarily to the setup of our exploration.

The analysis relies exclusively on offline evaluation, which constrains our ability to capture long-term behavioral adaptation and real-world user perception. Although offline metrics offer controlled comparability across models, they only provide an approximate of the actual effectiveness in deployment. Consequently, the impact of explicitly incorporating value information into a recommender system remains unknown outside of this controlled setting.

A further limitation concerns the extraction and representation of personal values. This study was constrained by the absence of ground truth labels for user values; to the best of our knowledge, no publicly available dataset includes explicit personal value annotations. We therefore relied on lexical analysis to derive value signals from textual content. Although prior work supports this as a valid and robust approach [17, 27], dependence on textual sources introduces noise. In particular, value profiles were inferred from book descriptions, and while most items in the dataset have such descriptions, a substantial portion does not, potentially reducing the accuracy of the inferred profiles for those items.

Our reliance on the PVD also comes with certain limitations. While this dictionary has been empirically proven to correspond with self-reported scores on the SVS (the "gold standard" self-reported survey used in most psychological research into values), the correlation between these PVD scores and SVS scores is moderate for seven out of ten values. Both the strength of the correlation, and the fact that the researchers did not find correspondence for the values Security, Conformity, and Power [169], might have influenced our findings.

The text preprocessing pipeline presents a potential area for improvement. During tokenization, non-alphanumeric characters were removed, which in some cases caused adjacent tokens to be concatenated erroneously. As the value acquisition relies on exact lexical matching against the dictionary, such concatenations result in valid value-related words failing to be detected. To assess the impact of this preprocessing artifact, a targeted experiment was conducted comparing UserKNN and UserKNNUVP under corrected and uncorrected preprocessing conditions. The results indicate that this issue does not substantially affect model performance. Nevertheless, we do acknowledge that using the corrected preprocessing condition might results in a more complete view of the value hierarchy present in the text.

Connected to this, our method relied on exact matching, since this is what the PVD was built upon. However, when writing reviews, users might make spelling mistakes or use alternative spellings for

words, and in such cases exact matching would not return any results. While exact matching mirrors the method used when empirically testing the PVD, this might have limited the results.

Finally, the novelty and diversity metrics employed in this study rely primarily on genre information. While prior work has shown that users associate diversity and novelty with genre information [53, 218, 168], it may be insufficient to fully capture user perceptions of those concepts. As noted in related work, richer feature representations, such as plot summaries combined with genre labels [102], could result in more nuanced measurements.

8

Conclusion

This thesis provides insights into how user values are reflected in recommendation outcomes and whether incorporating explicit value profiles improves the alignment between users and recommender systems. Specifically, we investigate the following research questions:

- **RQ1:** To what extent do recommendations generated by a standard recommender system align with users' personal values?
- **RQ2:** Does incorporating user values as features within a recommender increase this alignment?

Addressing **RQ1**, we observe that across all evaluated models recommendations exhibit a weak degree of alignment with user value structures in the baseline setting. This indicates that value signals are implicitly present in interaction-driven recommendation data, even without explicit value modeling.

An important finding of our exploration considering **RQ2** is that introducing user value profiles as additional input significantly increases alignment across models. While the magnitude of improvement varies between models, the direction of effect is stable: introducing the UVP shifts recommendation outputs closer to users' value orientations. At the same time, this improvement is not uniform across all users, suggesting that value integration interacts with both model architecture and individual user profiles.

The increase in value alignment also comes with a series of trade-offs. The increased alignment is associated with decreased accuracy, and with mixed effects on the other metrics. The results highlight that creating value-aligned recommender systems is possible, but comes with certain caveats that should be taken into account. In particular, the trade-offs between value-alignment and other dimensions of recommendations need to be addressed. Beyond these considerations, we conclude that investigating value alignment is an effective way to increase autonomy in RSs, and encourage researchers to further investigate methods for early integration of value information.

8.1. Future Work

While the current investigation provides important insights into how personal values are reflected in recommendation outcomes, several questions remain unanswered at present. Future studies on the current topic are therefore recommended in the following areas.

Future work could investigate online evaluation methodologies to better assess the impact of value-aware recommendations. For example, it would be interesting to investigate whether users notice the effect of the increase in value-alignment. Related work has indicated that users are able to distinguish between non-aligned and aligned recommendations [97], but that work focuses on reranking. More research is needed to investigate whether this observation also holds for the early integration methods that we explored.

Furthermore, the experiment could be replicated in an altered for, using the SVS instead of implicit value acquisition. Certain work in this field has been done by [83], but those authors did not investigate

ranking metrics, nor did they ask users whether they thought that the recommendations were more value aligned.

Future research could improve value modeling by incorporating full-text analysis of book content or author-level value inference. Additionally, improving robustness in value extraction for difficult dimensions such as Power, Security, and Conformity remains an open challenge, as these constructs are less reliably captured in textual proxies.

A natural progression of this exploration would be to investigate value-alignment in other domains. For example, future work could investigate the impact of value-alignment in domains where the role of personal values might be greater, such as online dating. Besides being a natural fit for value-alignment, such an exploration would also extend knowledge of the effect of applying these methods in a recommendation technique, namely reciprocal recommender systems [209].

More broadly, research into integrating personal values into RSs would benefit from a dataset that includes ground truth personal values information, similar to personality datasets such as myPersonality [198], and Personality2018 [155]. These datasets have helped increase the effectiveness of personality-aware RSs, and we expect creating these datasets would cause similar improvements for research into value-aligned recommender systems.

Bibliography

[1] Panagiotis Adamopoulos and Alexander Tuzhilin. 2014. On over-specialization and concentration bias of recommendations: probabilistic neighborhood selection in collaborative filtering systems. In *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, Foster City, Silicon Valley California USA, 153–160. https://doi.org/10.1145/2645710.2645752

[2] Panagiotis Adamopoulos and Alexander Tuzhilin. 2014. On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 4 (2014), 1–32. Publisher: ACM New York, NY, USA.

[3] Gediminas Adomavicius, Konstantin Bauman, Alexander Tuzhilin, Moshe Unger, Lior Rokach, Francesco Ricci, and Bracha Shapira. 2022. Context-Aware Recommender Systems: From Foundations to Recent Developments. In *Recommender Systems Handbook*. Springer US, New York, NY, 211–250.

[4] Gediminas Adomavicius, Zan Huang, and Alexander Tuzhilin. 2008. Personalization and Recommender Systems. In *State-of-the-Art Decision-Making Tools in the Information-Intensive Age*, Zhi-Long Chen, S. Raghavan, Paul Gray, and Harvey J. Greenberg (Eds.). INFORMS, 55–107. https://doi.org/10.1287/educ.1080.0044

[5] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering* 17, 6 (2005), 734–749.

[6] Sofia Aftab and Heri Ramampiaro. 2022. Evaluating Top-N Recommendations Using Ranked Error Approach: An Empirical Analysis. *IEEE Access* 10 (2022), 30832–30845. https://doi.org/10.1109/ACCESS.2022.3159646

[7] Charu C Aggarwal. 2016. *Recommender Systems*. Vol. 1. Springer.

[8] Charu C Aggarwal and Tarek Abdelzaher. 2012. Social sensing. In *Managing and mining sensor data*. Springer, 237–297.

[9] Haldun Akoglu. 2018. User's guide to correlation coefficients. *Turkish journal of emergency medicine* 18, 3 (2018), 91–93.

[10] Bushra Alhijawi, Arafat Awajan, and Salam Fraihat. 2023. Survey on the Objectives of Recommender Systems: Measures, Solutions, Evaluation Methodology, and New Perspectives. *Comput. Surveys* 55, 5 (May 2023), 1–38. https://doi.org/10.1145/3527449

[11] Hind I. Alshbanat, Hafida Benhidour, and Said Kerrache. 2025. A survey of latent factor models in recommender systems. *Information Fusion* 117 (May 2025), 102905. https://doi.org/10.1016/j.inffus.2024.102905

[12] Xavier Amatriain, Josep M. Pujol, and Nuria Oliver. 2009. I Like It... I Like It Not: Evaluating User Ratings Noise in Recommender Systems. In *User Modeling, Adaptation, and Personalization*, Geert-Jan Houben, Gord McCalla, Fabio Pianesi, and Massimo Zancanaro (Eds.). Vol. 5535. Springer Berlin Heidelberg, Berlin, Heidelberg, 247–258. https://doi.org/10.1007/978-3-642-02247-0_24 Series Title: Lecture Notes in Computer Science.

[13] Vito Walter Anelli, Alejandro Bellogin, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021. Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Virtual Event Canada, 2405–2414. https://doi.org/10.1145/3404835.3463245

[14] Vito Walter Anelli, Yashar Deldjoo, Tommaso DiNoia, Felice Antonio Merra, Bracha Shapira, Lior Rokach, and Francesco Ricci. 2022. Adversarial Recommender Systems: Attack, Defense, and Advances. In *Recommender Systems Handbook*. Springer US, New York, NY, 335–379.

[15] Anthony J Applegate and Mary Dekonty Applegate. 2004. The Peter Effect: Reading Habits and Attitudes of Preservice Teachers. *The reading teacher* 57, 6 (2004), 554–563.

[16] Nana Yaw Asabere, Amevi Acakpovi, and Mathias Bennet Michael. 2018. Improving Socially-Aware Recommendation Accuracy Through Personality. *IEEE Transactions on Affective Computing* 9, 3 (July 2018), 351–361. https://doi.org/10.1109/TAFFC.2017.2695605

[17] Anat Bardi, Rachel M. Calogero, and Brian Mullen. 2008. A new archival approach to the study of values and value–Behavior relations: Validation of the value lexicon. *Journal of Applied Psychology* 93, 3 (2008), 483–497. https://doi.org/10.1037/0021-9010.93.3.483

[18] Christine Bauer, Chandni Bagchi, Olusanmi A. Hundogan, and Karin Van Es. 2024. Where Are the Values? A Systematic Literature Review on News Recommender Systems. *ACM Transactions on Recommender Systems* 2, 3 (Sept. 2024), 1–40. https://doi.org/10.1145/3654805

[19] Christine Bauer, Eva Zangerle, and Alan Said. 2024. Exploring the Landscape of Recommender Systems Evaluation: Practices and Perspectives. *ACM Transactions on Recommender Systems* 2, 1 (March 2024), 1–31. https://doi.org/10.1145/3629170

[20] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries* 17, 4 (Nov. 2016), 305–338. https://doi.org/10.1007/s00799-015-0156-0

[21] Alejandro Bellogín and Alan Said. 2021. Improving accountability in recommender systems research through reproducibility. *User Modeling and User-Adapted Interaction* 31, 5 (Nov. 2021), 941–977. https://doi.org/10.1007/s11257-021-09302-x

[22] Melanie Berger, Guillaume Pottier, Bastian Pfleging, and Regina Bernhaupt. 2022. Considering Users' Personal Values in User-Centered Design Processes for Media and Entertainment Services. In *Human-Centered Software Engineering*, Regina Bernhaupt, Carmelo Ardito, and Stefan Sauer (Eds.). Vol. 13482. Springer International Publishing, Cham, 129–139. https://doi.org/10.1007/978-3-031-14785-2_8 Series Title: Lecture Notes in Computer Science.

[23] Michael Bernstein, Angèle Christin, Jeffrey Hancock, Tatsunori Hashimoto, Chenyan Jia, Michelle Lam, Nicole Meister, Nathaniel Persily, Tiziano Piccardi, Martin Saveski, Jeanne Tsai, Johan Ugander, and Chunchen Xu. 2023. Embedding Societal Values into Social Media Algorithms. *Journal of Online Trust and Safety* 2, 1 (Sept. 2023). https://doi.org/10.54501/jots.v2i1.148

[24] Ingvild H. Blomstervik and Svein Ottar Olsen. 2024. The relationship between personal values and preference for novelty: conceptual issues and the novelty–familiarity continuum. *Current Issues in Tourism* (Nov. 2024), 1–18. https://doi.org/10.1080/13683500.2024.2428767

[25] Tesfaye Fenta Boka, Zhendong Niu, and Rama Bastola Neupane. 2024. A survey of sequential recommendation systems: Techniques, evaluation, and future directions. *Information Systems* 125 (Nov. 2024), 102427. https://doi.org/10.1016/j.is.2024.102427

[26] Sofia Bonicalzi, Mario De Caro, and Benedetta Giovanola. 2023. Artificial Intelligence and Autonomy: On the Ethical Dimension of Recommender Systems. *Topoi* 42, 3 (July 2023), 819–832. https://doi.org/10.1007/s11245-023-09922-5

[27] Ryan Boyd, Steven Wilson, James Pennebaker, Michal Kosinski, David Stillwell, and Rada Mihalcea. 2021. Values in Words: Using Language to Evaluate and Understand Personal Values. *Proceedings of the International AAAI Conference on Web and Social Media* 9, 1 (Aug. 2021), 31–40. https://doi.org/10.1609/icwsm.v9i1.14589

[28] Philip Brey. 2015. Design for the Value of Human Well-Being. In *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, Jeroen Van Den Hoven, Pieter E. Vermaas, and Ibo Van De Poel (Eds.). Springer Netherlands, Dordrecht. https://link.springer.com/10.1007/978-94-007-6970-0

[29] Robin Burke. 2002. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction* 12, 4 (2002), 331–370.

[30] Robin Burke. 2007. Hybrid web recommender systems. *The adaptive web: methods and strategies of web personalization* (2007), 377–408.

[31] Christopher Burr, Nello Cristianini, and James Ladyman. 2018. An Analysis of the Interaction Between Intelligent Software Agents and Human Users. *Minds and Machines* 28, 4 (Dec. 2018), 735–774. https://doi.org/10.1007/s11023-018-9479-0

[32] Rafael A. Calvo, Dorian Peters, Karina Vold, and Richard M. Ryan. 2020. Supporting Human Autonomy in AI Systems: A Framework for Ethical Enquiry. In *Ethics of Digital Well-Being*, Christopher Burr and Luciano Floridi (Eds.). Vol. 140. Springer International Publishing, Cham, 31–54. https://doi.org/10.1007/978-3-030-50585-1_2 Series Title: Philosophical Studies Series.

[33] Pablo Castells, Neil Hurley, and Saúl Vargas. 2022. Novelty and Diversity in Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, New York, NY, 603–646. https://doi.org/10.1007/978-1-0716-2197-4_16

[34] Rocío Cañamares, Pablo Castells, and Alistair Moffat. 2020. Offline evaluation options for recommender systems. *Information Retrieval Journal* 23, 4 (Aug. 2020), 387–410. https://doi.org/10.1007/s10791-020-09371-3

[35] Lidia Ceriani and Paolo Verme. 2012. The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *The Journal of Economic Inequality* 10, 3 (2012), 421–443.

[36] Jinpeng Chen, Huachen Guan, Hongbo Gao, Huan Li, Zhenye Yang, Fan Zhang, Kaimin Wei, Feifei Kou, and Xindong Wu. 2026. Enhancing Explainable Sequential Recommendation With Disentangled Representations and Auxiliary Review Explanations. *IEEE Transactions on Computational Social Systems* 13, 1 (Feb. 2026), 783–798. https://doi.org/10.1109/TCSS.2025.3603087

[37] Jilin Chen, Gary Hsieh, Jalal U. Mahmud, and Jeffrey Nichols. 2014. Understanding individuals' personal values from social media word use. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, Baltimore Maryland USA, 405–414. https://doi.org/10.1145/2531602.2531608

[38] Liqiong Chen, Xiaoyu Bi, Guoqing Fan, and Huaiying Sun. 2023. A multitask recommendation algorithm based on DeepFM and Graph Convolutional Network. *Concurrency and Computation: Practice and Experience* 35, 2 (Jan. 2023), e7498. https://doi.org/10.1002/cpe.7498

[39] Zhilong Chen, Jinghua Piao, Xiaochong Lan, Hancheng Cao, Chen Gao, Zhicong Lu, and Yong Li. 2022. Practitioners Versus Users: A Value-Sensitive Evaluation of Current Industrial Recommender System Design. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–32. https://doi.org/10.1145/3555646

[40] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & Deep Learning for Recommender Systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*. ACM, Boston MA USA, 7–10. https://doi.org/10.1145/2988450.2988454

[41] Seunghwan Choi, Donghoon Lee, Hyeoungguk Kang, and Hyunsouk Cho. 2025. Exploring the Side-Information Fusion for Sequential Recommendation. *IEEE Access* 13 (2025), 8839–8850. https://doi.org/10.1109/ACCESS.2025.3525812

[42] Markus Christen, Darcia Narvaez, Carmen Tanner, and Thomas Ott. 2016. Using thesauruses as a heuristics for mapping values. *Cognitive Systems Research* 40 (Dec. 2016), 59–74. https://doi.org/10.1016/j.cogsys.2016.02.003

[43] Cindy K Chung, Peter J Rentfrow, and James W Pennebaker. 2014. Finding values in words: Using natural language to detect regional variations in personal concerns. In *Geographical psychology: Exploring the interaction of environment and behavior*. American Psychological Association, 195–216. https://doi.org/10.1037/14272-011

[44] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

[45] Andrew Collins, Dominika Tkaczyk, Akiko Aizawa, and Joeran Beel. 2018. Position Bias in Recommender Systems for Digital Libraries. In *Transforming Digital Worlds*, Gobinda Chowdhury, Julie McLeod, Val Gillet, and Peter Willett (Eds.). Vol. 10766. Springer International Publishing, Cham, 335–344. https://doi.org/10.1007/978-3-319-78105-1_37 Series Title: Lecture Notes in Computer Science.

[46] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, Barcelona Spain, 39–46. https://doi.org/10.1145/1864708.1864721

[47] Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological bulletin* 52, 4 (1955), 281.

[48] Savvina Daniil, Mirjam Cuper, Cynthia C. S. Liem, Jacco Van Ossenbruggen, and Laura Hollink. 2024. Reproducing Popularity Bias in Recommendation: The Effect of Evaluation Strategies. *ACM Transactions on Recommender Systems* 2, 1 (March 2024), 1–39. https://doi.org/10.1145/3637066

[49] Abhinandan S. Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*. ACM, Banff Alberta Canada, 271–280. https://doi.org/10.1145/1242572.1242610

[50] Juan Ignacio Del Valle and Francisco Lara. 2024. AI-powered recommender systems and the preservation of personal autonomy. *AI & SOCIETY* 39, 5 (Oct. 2024), 2479–2491. https://doi.org/10.1007/s00146-023-01720-2

[51] P. Devika and A. Milton. 2024. Book recommendation system: reviewing different techniques and approaches. *International Journal on Digital Libraries* 25, 4 (Dec. 2024), 803–824. https://doi.org/10.1007/s00799-024-00403-7

[52] Juhi Dhameliya and Nikita Desai. 2019. Job Recommender Systems: A Survey. In *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, Vol. 1. 1–5. https://doi.org/10.1109/i-PACT44901.2019.8960231

[53] Patrik Dokoupil, Ludovico Boratto, and Ladislav Peska. 2024. User Perceptions of Diversity in Recommender Systems. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. ACM, Cagliari Italy, 212–222. https://doi.org/10.1145/3627043.3659555

[54] Patrik Dokoupil, Ludovico Boratto, and Ladislav Peska. 2025. How Do Users Perceive Recommender Systems' Objectives?. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. ACM, Prague Czech Republic, 165–176. https://doi.org/10.1145/3705328.3748066

[55] Sabina Eftimova, Magdalena Garvanova, and B. Nikolova. 2021. Reading as a key factor for personality development of adolescents. In *EDULEARN21 proceedings*. Online Conference, 3280–3286. https://doi.org/10.21125/edulearn.2021.0696

[56] Michael D. Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz, Bracha Shapira, Lior Rokach, and Francesco Ricci. 2022. Fairness in Recommender Systems. In *Recommender Systems Handbook*. Springer US, New York, NY, 679–707.

[57] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. 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 (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 172–186. https://proceedings.mlr.press/v81/ekstrand18b.html

[58] Mehdi Elahi, Matthias Braunhofer, Francesco Ricci, and Marko Tkalcic. 2013. Personality-Based Active Learning for Collaborative Filtering Recommender Systems. In *AI*IA 2013: Advances in Artificial Intelligence*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Matteo Baldoni, Cristina Baroglio, Guido Boella, and Roberto Micalizio (Eds.). Vol. 8249. Springer International Publishing, Cham, 360–371. https://doi.org/10.1007/978-3-319-03524-6_31 Series Title: Lecture Notes in Computer Science.

[59] Farzad Eskandanian, Bamshad Mobasher, and Robin Burke. 2017. A Clustering Approach for Personalizing Diversity in Collaborative Recommender Systems. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. ACM, Bratislava Slovakia, 280–284. https://doi.org/10.1145/3079628.3079699

[60] Matteo Fabbri. 2023. Social influence for societal interest: a pro-ethical framework for improving human decision making through multi-stakeholder recommender systems. *AI & SOCIETY* 38, 2 (April 2023), 995–1002. https://doi.org/10.1007/s00146-022-01467-2

[61] Norman T Feather. 1995. Values, valences, and choice: The influences of values on the perceived attractiveness and choice of alternatives. *Journal of personality and social psychology* 68, 6 (1995), 1135.

[62] Ignacio Fernández-Tobías, Matthias Braunhofer, Mehdi Elahi, Francesco Ricci, and Iván Cantador. 2016. Alleviating the new user problem in collaborative filtering by exploiting personality information. *User Modeling and User-Adapted Interaction* 26, 2-3 (June 2016), 221–255. https://doi.org/10.1007/s11257-016-9172-z

[63] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research. *ACM Transactions on Information Systems* 39, 2 (April 2021), 1–49. https://doi.org/10.1145/3434185

[64] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, Copenhagen Denmark, 101–109. https://doi.org/10.1145/3298689.3347058

[65] Lior Fink, Leorre Newman, and Uriel Haran. 2024. Let me decide: Increasing user autonomy increases recommendation acceptance. *Computers in Human Behavior* 156 (July 2024), 108244. https://doi.org/10.1016/j.chb.2024.108244

[66] Johnny RJ Fontaine, Ype H Poortinga, Luc Delbeke, and Shalom H Schwartz. 2008. Structural equivalence of the values domain across cultures: Distinguishing sampling fluctuations from meaningful variation. *Journal of Cross-Cultural Psychology* 39, 4 (2008), 345–365.

[67] Reinaldo Silva Fortes, Daniel Xavier De Sousa, Dayanne G. Coelho, Anisio M. Lacerda, and Marcos A. Gonçalves. 2021. Individualized extreme dominance (IndED): A new preference-based method for multi-objective recommender systems. *Information Sciences* 572 (Sept. 2021), 558–573. https://doi.org/10.1016/j.ins.2021.05.037

[68] Reinaldo Silva Fortes, Anisio Lacerda, Alan Freitas, Carlos Bruckner, Dayanne Coelho, and Marcos Gonçalves. 2018. User-Oriented Objective Prioritization for Meta-Featured Multi-Objective Recommender Systems. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, Singapore Singapore, 311–316. https://doi.org/10.1145/3213586.3225243

[69] Batya Friedman, Peter H. Kahn, Alan Borning, and Alina Huldtgren. 2013. Value Sensitive Design and Information Systems. In *Early engagement and new technologies: Opening up the laboratory*, Neelke Doorn, Daan Schuurbiers, Ibo van de Poel, and Michael E. Gorman (Eds.). Springer Netherlands, Dordrecht, 55–95. https://doi.org/10.1007/978-94-007-7844-3_4

[70] R Michael Furr. 2008. A framework for profile similarity: Integrating similarity, normativeness, and distinctiveness. *Journal of personality* 76, 5 (2008), 1267–1316.

[71] Chongming Gao, Ruijun Chen, Shuai Yuan, Kexin Huang, Yuanqing Yu, and Xiangnan He. 2025. SPRec: Self-Play to Debias LLM-based Recommendation. In *Proceedings of the ACM on Web Conference 2025*. ACM, Sydney NSW Australia, 5075–5084. https://doi.org/10.1145/3696410.3714524

[72] Angel L. Garrido, Maria Soledad Pera, and Sergio Ilarri. 2014. SOLE-R: A Semantic and Linguistic Approach for Book Recommendations. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*. IEEE, Athens, Greece, 524–528. https://doi.org/10.1109/ICALT.2014.155

[73] Damianos Gavalas, Charalampos Konstantopoulos, Konstantinos Mastakas, and Grammati Pantziou. 2014. Mobile recommender systems in tourism. *Journal of Network and Computer Applications* 39 (March 2014), 319–333. https://doi.org/10.1016/j.jnca.2013.04.006

[74] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, Barcelona Spain, 257–260. https://doi.org/10.1145/1864708.1864761

[75] Francesco Gelli, Xiangnan He, Tao Chen, and Tat-Seng Chua. 2017. How Personality Affects our Likes: Towards a Better Understanding of Actionable Images. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, Mountain View California USA, 1828–1837. https://doi.org/10.1145/3123266.3127909

[76] Bingrui Geng, Lingling Li, Licheng Jiao, Maoguo Gong, Qing Cai, and Yue Wu. 2015. NNIA-RS: A multi-objective optimization based recommender system. *Physica A: Statistical Mechanics and its Applications* 424 (April 2015), 383–397. https://doi.org/10.1016/j.physa.2015.01.007

[77] Sergio Genovesi, Katharina Kaesling, and Scott Robbins (Eds.). 2023. *Recommender Systems: Legal and Ethical Issues*. The International Library of Ethics, Law and Technology, Vol. 40. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-34804-4

[78] Mustansar Ali Ghazanfar. 2015. Experimenting switching hybrid recommender systems. *Intelligent Data Analysis* 19, 4 (2015), 845–877.

[79] Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. 2016. What does research reproducibility mean? *Science Translational Medicine* 8, 341 (June 2016). https://doi.org/10.1126/scitranslmed.aaf5027

[80] Robin Goodwin and Merlin Tinker. 2002. Value priorities and preferences for a relationship partner. *Personality and Individual Differences* 32, 8 (2002), 1339–1349.

[81] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, 1725–1731. https://doi.org/10.24963/ijcai.2017/239

[82] Javier Guzmán-Obando, Josep Lluís De La Rosa, Silvana Aciar, Miquel Montaner, José A. Castán, and Julio Laria. 2008. The User's Human Values Scale Methodology in Recommender Systems from Several Information Sources of the Organization. In *MICAI 2008: Advances in Artificial Intelligence*, Alexander Gelbukh and Eduardo F. Morales (Eds.). Vol. 5317. Springer Berlin Heidelberg, Berlin, Heidelberg, 900–912. https://doi.org/10.1007/978-3-540-88636-5_85 Series Title: Lecture Notes in Computer Science.

[83] Mohammad Hajarian, Miguel Herrera Carrillo, Paloma Díaz, and Ignacio Aedo. 2025. Gamispotify: a gamified social music recommendation system based on users' personal values. *Multimedia Tools and Applications* (Feb. 2025). https://doi.org/10.1007/s11042-024-20588-y

[84] Jaron Harambam, Dimitrios Bountouridis, Mykola Makhortykh, and Joris Van Hoboken. 2019. Designing for the better by taking users into account: a qualitative evaluation of user control mechanisms in (news) recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, Copenhagen Denmark, 69–77. https://doi.org/10.1145/3298689.3347014

[85] aAdli Ihsan Hariadi and Dade Nurjanah. 2017. Hybrid attribute and personality based recommender system for book recommendation. In *2017 International Conference on Data and Software Engineering (ICoDSE)*. IEEE, Palembang, 1–5. https://doi.org/10.1109/ICODSE.2017.8285874

[86] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[87] Yun He, Jianling Wang, Wei Niu, and James Caverlee. 2019. A Hierarchical Self-Attentive Model for Recommending User-Generated Item Lists. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. ACM, Beijing China, 1481–1490. https://doi.org/10.1145/3357384.3358030

[88] Yun He, Yin Zhang, Weiwen Liu, and James Caverlee. 2020. Consistency-Aware Recommendation for User-Generated Item List Continuation. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. ACM, Houston TX USA, 250–258. https://doi.org/10.1145/3336191.3371776

[89] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems* 22, 1 (Jan. 2004), 5–53. https://doi.org/10.1145/963770.963772

[90] Steven Hitlin. 2003. Values as the Core of Personal Identity: Drawing Links between Two Theories of Self. *Social Psychology Quarterly* 66, 2 (June 2003), 118. https://doi.org/10.2307/1519843

[91] Steven Hitlin and Jane Allyn Piliavin. 2004. Values: Reviving a Dormant Concept. *Annual Review of Sociology* 30, 1 (Aug. 2004), 359–393. https://doi.org/10.1146/annurev.soc.30.012703.110640

[92] David Holtz, Ben Carterette, Praveen Chandar, Zahra Nazari, Henriette Cramer, and Sinan Aral. 2020. The Engagement-Diversity Connection: Evidence from a Field Experiment on Spotify. In *Proceedings of the 21st ACM Conference on Economics and Computation*. ACM, Virtual Event Hungary, 75–76. https://doi.org/10.1145/3391403.3399532

[93] Gary Hsieh, Jilin Chen, Jalal U. Mahmud, and Jeffrey Nichols. 2014. You read what you value: understanding personal values and reading interests. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Toronto Ontario Canada, 983–986. https://doi.org/10.1145/2556288.2556995

[94] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*. Ieee, 263–272.

[95] Ran Huang. 2023. Improved content recommendation algorithm integrating semantic information. *Journal of Big Data* 10, 1 (May 2023), 84. https://doi.org/10.1186/s40537-023-00776-7

[96] Yinghui Huang, Yuhang Dong, Weiqing Li, and Yue Xu. 2025. Can psychographics mitigate over-specialization in recommender-driven consumer markets? Evidence from recommender systems based simulation experiment. *Management System Engineering* 4, 1 (Nov. 2025), 20. https://doi.org/10.1007/s44176-025-00054-1

[97] Farnaz Jahanbakhsh, Dora Zhao, Tiziano Piccardi, Zachary Robertson, Ziv Epstein, Sanmi Koyejo, and Michael S. Bernstein. 2026. Value Alignment of Social Media Ranking Algorithms. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM, Barcelona Spain, 1–26. https://doi.org/10.1145/3772318.3791281

[98] Dietmar Jannach and Gediminas Adomavicius. 2016. Recommendations with a Purpose. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, Boston Massachusetts USA, 7–10. https://doi.org/10.1145/2959100.2959186

[99] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction* 25, 5 (Dec. 2015), 427–491. https://doi.org/10.1007/s11257-015-9165-3

[100] Dietmar Jannach, Bamshad Mobasher, and Shlomo Berkovsky. 2020. Research directions in session-based and sequential recommendation: A preface to the special issue. *User Modeling and User-Adapted Interaction* 30, 4 (Sept. 2020), 609–616. https://doi.org/10.1007/s11257-020-09274-4

[101] Dietmar Jannach, Massimo Quadrana, Paolo Cremonesi, Bracha Shapira, Lior Rokach, and Francesco Ricci. 2022. Session-Based Recommender Systems. In *Recommender Systems Handbook*. Springer US, New York, NY, 301–334.

[102] Mathias Jesse, Christine Bauer, and Dietmar Jannach. 2023. Intra-list similarity and human diversity perceptions of recommendations: the details matter. *User Modeling and User-Adapted Interaction* 33, 4 (Sept. 2023), 769–802. https://doi.org/10.1007/s11257-022-09351-w

[103] Mathias Jesse and Dietmar Jannach. 2021. Digital nudging with recommender systems: Survey and future directions. *Computers in Human Behavior Reports* 3 (Jan. 2021), 100052. https://doi.org/10.1016/j.chbr.2020.100052

[104] Mingi Ji, Weonyoung Joo, Kyungwoo Song, Yoon-Yeong Kim, and Il-Chul Moon. 2020. Sequential Recommendation with Relation-Aware Kernelized Self-Attention. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 04 (April 2020), 4304–4311. https://doi.org/10.1609/aaai.v34i04.5854

[105] Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. 2023. A Critical Study on Data Leakage in Recommender System Offline Evaluation. *ACM Transactions on Information Systems* 41, 3 (July 2023), 1–27. https://doi.org/10.1145/3569930

[106] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately Interpreting Clickthrough Data as Implicit Feedback. *SIGIR forum* 51, 1 (2017), 4–11.

[107] Michael Jugovac, Dietmar Jannach, and Lukas Lerche. 2017. Efficient optimization of multiple recommendation quality factors according to individual user tendencies. *Expert Systems with Applications* 81 (Sept. 2017), 321–331. https://doi.org/10.1016/j.eswa.2017.03.055

[108] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4 (Oct. 2002), 422–446. https://doi.org/10.1145/582415.582418

[109] Johannes Kaiser and Thorsten Quandt. 2016. Book lovers, bibliophiles, and fetishists: The social benefits of heavy book usage. *Psychology of Popular Media Culture* 5, 4 (Oct. 2016), 356–371. https://doi.org/10.1037/ppm0000077

[110] Marius Kaminskas and Derek Bridge. 2017. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Transactions on Interactive Intelligent Systems* 7, 1 (March 2017), 1–42. https://doi.org/10.1145/2926720

[111] Wang-Cheng Kang and Julian McAuley. 2018. Self-Attentive Sequential Recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, Singapore, 197–206. https://doi.org/10.1109/ICDM.2018.00035

[112] Komal Kapoor, Vikas Kumar, Loren Terveen, Joseph A. Konstan, and Paul Schrater. 2015. "I like to explore sometimes": Adapting to Dynamic User Novelty Preferences. In *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, Vienna Austria, 19–26. https://doi.org/10.1145/2792838.2800172

[113] Przemysław Kazienko and Erik Cambria. 2024. Toward Responsible Recommender Systems. *IEEE Intelligent Systems* 39, 3 (May 2024), 5–12. https://doi.org/10.1109/MIS.2024.3398190

[114] M. G. Kendall. 1945. The Treatment of Ties in Ranking Problems. *Biometrika* 33, 3 (1945), 239–251.

[115] Mehdi Khamassi, Marceau Nahon, and Raja Chatila. 2024. Strong and weak alignment of large language models with human values. *Scientific Reports* 14, 1 (Aug. 2024), 19399. https://doi.org/10.1038/s41598-024-70031-3

[116] Harry Khamis. 2008. Measures of Association: How to Choose? *Journal of Diagnostic Medical Sonography* 24, 3 (May 2008), 155–162. https://doi.org/10.1177/8756479308317006

[117] Euna Mehnaz Khan, Md. Saddam Hossain Mukta, Mohammed Eunus Ali, and Jalal Mahmud. 2020. Predicting Users' Movie Preference and Rating Behavior from Personality and Values. *ACM Transactions on Interactive Intelligent Systems* 10, 3 (Sept. 2020), 1–25. https://doi.org/10.1145/3338244

[118] Mohammed Khwaja, Miquel Ferrer, Jesus Omana Iglesias, A. Aldo Faisal, and Aleksandar Matic. 2019. Aligning daily activities with personality: towards a recommender system for improving well-being. In *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, Copenhagen Denmark, 368–372. https://doi.org/10.1145/3298689.3347020

[119] Hye-young Kim, Minjin Choi, Sunkyung Lee, Ilwoong Baek, and Jongwuk Lee. 2025. DIFF: Dual Side-Information Filtering and Fusion for Sequential Recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Padua Italy, 1624–1633. https://doi.org/10.1145/3726302.3729948

[120] Oliver Klingefjord, Ryan Lowe, and Joe Edelman. 2024. What are human values, and how do we align AI to them? https://doi.org/10.48550/arXiv.2404.10636 arXiv:2404.10636 [cs.CY].

[121] Yasamin Klingler, Claude Lehmann, João Pedro Monteiro, Carlo Saladin, Abraham Bernstein, and Kurt Stockinger. 2022. Evaluation of Algorithms for Interaction-Sparse Recommendations: Neural Networks don't Always Win. In *Proceedings of the 25th International Conference on Extending Database Technology, EDBT 2022, Edinburgh, UK, March 29 - April 1, 2022*, Julia Stoyanovich, Jens Teubner, Paolo Guagliardo, Milos Nikolic, Andreas Pieris, Jan Mühlig, Fatma Özcan, Sebastian Schelter, H. V. Jagadish, and Meihui Zhang (Eds.). OpenProceedings.org, 2:475–2:486. https://doi.org/10.48786/EDBT.2022.42

[122] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (Oct. 2012), 441–504. https://doi.org/10.1007/s11257-011-9118-4

[123] Yehuda Koren. 2009. Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 447–456.

[124] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[125] Yehuda Koren, Steffen Rendle, Robert Bell, Bracha Shapira, Lior Rokach, and Francesco Ricci. 2022. Advances in Collaborative Filtering. In *Recommender Systems Handbook*. Springer US, New York, NY, 91–142.

[126] Balázs Kovács. 2025. Five Is the Brightest Star. But by how Much? Testing the Equidistance of Star Ratings in Online Reviews. *Organizational Research Methods* 28, 2 (April 2025), 269–295. https://doi.org/10.1177/10944281231223412

[127] Walid Krichene and Steffen Rendle. 2020. On Sampled Metrics for Item Recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, Virtual Event CA USA, 1748–1757. https://doi.org/10.1145/3394486.3403226

[128] Joshua Krook and Jan Blockx. 2023. Recommender Systems, Autonomy and User Engagement. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*. ACM, Edinburgh United Kingdom, 1–9. https://doi.org/10.1145/3597512.3599712

[129] Joseph B Kruskal. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29, 1 (1964), 1–27.

[130] Athanasios Krystallis, Marco Vassallo, and George Chryssohoidis. 2012. The usefulness of Schwartz's 'Values Theory' in understanding consumer behaviour towards differentiated products. *Journal of Marketing Management* 28, 11-12 (Oct. 2012), 1438–1463. https://doi.org/10.1080/0267257X.2012.715091

[131] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in recommender systems – A survey. *Knowledge-Based Systems* 123 (May 2017), 154–162. https://doi.org/10.1016/j.knosys.2017.02.009

[132] Paul Kuyer and Bert Gordijn. 2023. Nudge in perspective: A systematic literature review on the ethical issues with nudging. *Rationality and Society* 35, 2 (May 2023), 191–230. https://doi.org/10.1177/10434631231155005

[133] Arto Laitinen and Otto Sahlgren. 2021. AI Systems and Respect for Human Autonomy. *Frontiers in Artificial Intelligence* 4 (Oct. 2021), 705164. https://doi.org/10.3389/frai.2021.705164

[134] Julie Anne Lee, Geoffrey N. Soutar, Timothy M. Daly, and Jordan J. Louviere. 2011. Schwartz Values Clusters in the United States and China. *Journal of Cross-Cultural Psychology* 42, 2 (March 2011), 234–252. https://doi.org/10.1177/0022022110396867

[135] Sointu Leikas, Ville-Juhani Ilmarinen, Markku Verkasalo, Hanna-Leena Vartiainen, and Jan-Erik Lönnqvist. 2018. Relationship satisfaction and similarity of personality traits, personal values, and attitudes. *Personality and Individual Differences* 123 (2018), 191–198.

[136] Jiayi Liao, Ruobing Xie, Sihang Li, Xiang Wang, Xingwu Sun, Zhanhui Kang, and Xiangnan He. 2025. Multi-Grained Patch Training for Efficient LLM-based Recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Padua Italy, 1572–1581. https://doi.org/10.1145/3726302.3730042

[137] Chang Liu, Xiaoguang Li, Guohao Cai, Zhenhua Dong, Hong Zhu, and Lifeng Shang. 2021. Non-invasive Self-attention for Side Information Fusion in Sequential Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 5 (May 2021), 4249–4256. https://doi.org/10.1609/aaai.v35i5.16549

[138] Qijiong Liu, Jiaren Xiao, Lu Fan, Jieming Zhu, and Xiao-Ming Wu. 2024. Learning Category Trees for ID-Based Recommendation: Exploring the Power of Differentiable Vector Quantization. In *Proceedings of the ACM Web Conference 2024*. ACM, Singapore Singapore, 3521–3532. https://doi.org/10.1145/3589334.3645484

[139] Qijiong Liu, Jieming Zhu, Jiahao Wu, Tiandeng Wu, Zhenhua Dong, and Xiao-Ming Wu. 2023. FANS: Fast Non-Autoregressive Sequence Generation for Item List Continuation. In *Proceedings of the ACM Web Conference 2023*. ACM, Austin TX USA, 3309–3318. https://doi.org/10.1145/3543507.3583430

[140] Babak Loni, Roberto Pagano, Martha Larson, and Alan Hanjalic. 2016. Bayesian Personalized Ranking with Multi-Channel User Feedback. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, Boston Massachusetts USA, 361–364. https://doi.org/10.1145/2959100.2959163

[141] Malte Ludewig, Noemi Mauro, Sara Latifi, and Dietmar Jannach. 2019. Performance comparison of neural and non-neural approaches to session-based recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 462–466. https://doi.org/10.1145/3298689.3347041

[142] Xin Luo, MengChu Zhou, Shuai Li, Zhuhong You, Yunni Xia, and Qingsheng Zhu. 2016. A Nonnegative Latent Factor Model for Large-Scale Sparse Matrices in Recommender Systems via Alternating Direction Method. *IEEE Transactions on Neural Networks and Learning Systems* 27, 3 (March 2016), 579–592. https://doi.org/10.1109/TNNLS.2015.2415257

[143] Chen Ma, Peng Kang, and Xue Liu. 2019. Hierarchical Gating Networks for Sequential Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, Anchorage AK USA, 825–833. https://doi.org/10.1145/3292500.3330984

[144] Suman Kalyan Maity, Abhishek Panigrahi, and Animesh Mukherjee. 2019. Analyzing Social Book Reading Behavior on Goodreads and How It Predicts Amazon Best Sellers. In *Influence and Behavior Analysis in Social Networks and Social Media*, Mehmet Kaya and Reda Alhajj (Eds.). Springer International Publishing, Cham, 211–235. https://doi.org/10.1007/978-3-030-02592-2_11

[145] Saranya Maneeroj and Nakarin Sritrakool. 2022. An End-to-End Personalized Preference Drift Aware Sequential Recommender System With Optimal Item Utilization. *IEEE Access* 10 (2022), 62932–62952. https://doi.org/10.1109/ACCESS.2022.3182390

[146] Olga V. Maslova, Dmitry A. Shlyakhta, and Mikhail S. Yanitskiy. 2020. Schwartz Value Clusters in Modern University Students. *Behavioral Sciences* 10, 3 (March 2020), 66. https://doi.org/10.3390/bs10030066

[147] Sean M McNee, John Riedl, and Joseph A Konstan. 2006. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*. 1097–1101.

[148] Zaiqiao Meng, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2020. Exploring Data Splitting Strategies for the Evaluation of Recommendation Models. In *Fourteenth ACM Conference on Recommender Systems*. ACM, Virtual Event Brazil, 681–686. https://doi.org/10.1145/3383313.3418479

[149] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender systems and their ethical challenges. *AI & SOCIETY* 35, 4 (Dec. 2020), 957–967. https://doi.org/10.1007/s00146-020-00950-y

[150] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2021. Ethical aspects of multi-stakeholder recommendation systems. *The Information Society* 37, 1 (Jan. 2021), 35–45. https://doi.org/10.1080/01972243.2020.1832636

[151] Stuart Mills and Henrik Skaug Sætra. 2024. The autonomous choice architect. *AI & SOCIETY* 39, 2 (April 2024), 583–595. https://doi.org/10.1007/s00146-022-01486-z

[152] Aleksandr Milogradskii, Oleg Lashinin, Alexander P, Marina Ananyeva, and Sergey Kolesnikov. 2024. Revisiting BPR: A Replicability Study of a Common Recommender System Baseline. In *18th ACM Conference on Recommender Systems*. ACM, Bari Italy, 267–277. https://doi.org/10.1145/3640457.3688073

[153] Davide Morselli, Dario Spini, and Thierry Devos. 2012. Human values and trust in institutions across countries: A multilevel test of Schwartz's hypothesis of structural equivalence. In *Survey Research Methods*, Vol. 6. 49–60.

[154] Cataldo Musto, Marco de Gemmis, Pasquale Lops, Fedelucio Narducci, Giovanni Semeraro, Bracha Shapira, Lior Rokach, and Francesco Ricci. 2022. Semantics and Content-Based Recommendations. In *Recommender Systems Handbook*. Springer US, New York, NY, 251–298.

[155] Tien T. Nguyen, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2018. User Personality and User Satisfaction with Recommender Systems. *Information Systems Frontiers* 20, 6 (Dec. 2018), 1173–1189. https://doi.org/10.1007/s10796-017-9782-y

[156] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 188–197. https://doi.org/10.18653/v1/D19-1018

[157] Athanasios N. Nikolakopoulos, Xia Ning, Christian Desrosiers, George Karypis, Bracha Shapira, Lior Rokach, and Francesco Ricci. 2022. Trust Your Neighbors: A Comprehensive Survey of Neighborhood-Based Methods for Recommender Systems. In *Recommender Systems Handbook*. Springer US, New York, NY, 39–89.

[158] Huansheng Ning, Sahraoui Dhelim, and Nyothiri Aung. 2019. PersoNet: Friend Recommendation System Based on Big-Five Personality Traits and Hybrid Filtering. *IEEE Transactions on Computational Social Systems* 6, 3 (June 2019), 394–402. https://doi.org/10.1109/TCSS.2019.2903857

[159] Umaporn Padungkiatwattana, Thitiya Sae-Diae, Saranya Maneeroj, and Atsuhiro Takasu. 2022. ARERec: Attentive Local Interaction Model for Sequential Recommendation. *IEEE Access* 10 (2022), 31340–31358. https://doi.org/10.1109/ACCESS.2022.3160466

[160] Li-Wei Pan, Wei-Ke Pan, Mei-Yan Wei, Hong-Zhi Yin, and Zhong Ming. 2026. A survey on sequential recommendation. *Frontiers of Computer Science* 20, 3 (March 2026), 2003606. https://doi.org/10.1007/s11704-025-41329-w

[161] Vincenzo Paparella, Dario Di Palma, Vito Walter Anelli, and Tommaso Di Noia. 2023. Broadening the Scope: Evaluating the Potential of Recommender Systems beyond prioritizing Accuracy. In *Proceedings of the 17th ACM Conference on Recommender Systems*. ACM, Singapore Singapore, 1139–1145. https://doi.org/10.1145/3604915.3610649

[162] Javier Parapar and Filip Radlinski. 2021. Towards Unified Metrics for Accuracy and Diversity for Recommender Systems. In *Fifteenth ACM Conference on Recommender Systems*. ACM, Amsterdam Netherlands, 75–84. https://doi.org/10.1145/3460231.3474234

[163] Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. In *The adaptive web: methods and strategies of web personalization*. Springer, 325–341.

[164] Bo Peng, Ziqi Chen, Srinivasan Parthasarathy, and Xia Ning. 2024. Modeling Sequences as Star Graphs to Address Over-Smoothing in Self-Attentive Sequential Recommendation. *ACM Transactions on Knowledge Discovery from Data* 18, 8 (Sept. 2024), 1–24. https://doi.org/10.1145/3676560

[165] Maria Soledad Pera and Yiu Kai Ng. 2014. How Can We Help Our K-12 Teachers?: Using a Recommender to Make Personalized Book Suggestions. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. IEEE, Warsaw, Poland, 335–342. https://doi.org/10.1109/WI-IAT.2014.116

[166] Maria Soledad Pera and Yiu-Kai Ng. 2015. Analyzing Book-Related Features to Recommend Books for Emergent Readers. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*. ACM Press, Guzelyurt, Northern Cyprus, 221–230. https://doi.org/10.1145/2700171.2791037

[167] Dorian Peters and Rafael A. Calvo. 2023. Self-Determination Theory and Technology Design. In *The Oxford Handbook of Self-Determination Theory* (1 ed.), Richard M. Ryan (Ed.). Oxford University Press, 978–999. https://doi.org/10.1093/oxfordhb/9780197600047.013.49

[168] Yanni Ping, Yang Li, and Jiaxin Zhu. 2025. Beyond accuracy measures: the effect of diversity, novelty and serendipity in recommender systems on user engagement. *Electronic Commerce Research* 25, 3 (June 2025), 2177–2204. https://doi.org/10.1007/s10660-024-09813-w

[169] Vladimir Ponizovskiy, Murat Ardag, Lusine Grigoryan, Ryan Boyd, Henrik Dobewall, and Peter Holtz. 2020. Development and Validation of the Personal Values Dictionary: A Theory–Driven Tool for Investigating References to Basic Human Values in Text. *European Journal of Personality* 34, 5 (Sept. 2020), 885–902. https://doi.org/10.1002/per.2294

[170] Pearl Pu and Li Chen. 2010. A User-Centric Evaluation Framework of Recommender Systems. 612 (2010).

[171] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. 2019. Sequence-Aware Recommender Systems. *Comput. Surveys* 51, 4 (July 2019), 1–36. https://doi.org/10.1145/3190616

[172] Devangam Bangaru Rajesh and Avadhesh Kumar. 2025. Collaborative filtering models an experimental and detailed comparative study. *Scientific Reports* 15, 1 (2025), 31667.

[173] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09)*. AUAI Press, Arlington, Virginia, USA, 452–461.

[174] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized Markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. ACM, Raleigh North Carolina USA, 811–820. https://doi.org/10.1145/1772690.1772773

[175] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work - CSCW '94*. ACM Press, Chapel Hill, North Carolina, United States, 175–186. https://doi.org/10.1145/192844.192905

[176] Wondo Rhee, Sung Min Cho, and Bongwon Suh. 2022. Countering Popularity Bias by Regularizing Score Differences. In *Proceedings of the 16th ACM Conference on Recommender Systems*. ACM, Seattle WA USA, 145–155. https://doi.org/10.1145/3523227.3546757

[177] Marco Tulio Ribeiro, Nivio Ziviani, Edleno Silva De Moura, Itamar Hata, Anisio Lacerda, and Adriano Veloso. 2015. Multiobjective Pareto-Efficient Approaches for Recommender Systems. *ACM Transactions on Intelligent Systems and Technology* 5, 4 (Jan. 2015), 1–20. https://doi.org/10.1145/2629350

[178] Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). 2022. *Recommender Systems Handbook*. Springer US, New York, NY. https://doi.org/10.1007/978-1-0716-2197-4

[179] Marko A. Rodriguez and Jennifer H. Watkins. 2009. Faith in the Algorithm, Part 2: Computational Eudaemonics. In *Knowledge-Based and Intelligent Information and Engineering Systems*, Juan D. Velásquez, Sebastián A. Ríos, Robert J. Howlett, and Lakhmi C. Jain (Eds.). Vol. 5712. Springer Berlin Heidelberg, Berlin, Heidelberg, 813–820. https://doi.org/10.1007/978-3-642-04592-9_101 Series Title: Lecture Notes in Computer Science.

[180] Katherine H. Rogers, Dustin Wood, and R. Michael Furr. 2018. Assessment of similarity and self-other agreement in dyadic relationships: A guide to best practices. *Journal of Social and Personal Relationships* 35, 1 (Jan. 2018), 112–134. https://doi.org/10.1177/0265407517712615

[181] Maksim Rudnev. 2021. Caveats of non-ipsatization of basic values: A review of issues and a simulation study. *Journal of Research in Personality* 93 (Aug. 2021), 104118. https://doi.org/10.1016/j.jrp.2021.104118

[182] Willem E Saris, Desiree Knoppen, and Shalom H Schwartz. 2013. Operationalizing the theory of human values: Balancing homogeneity of reflective items and theoretical coverage. In *Survey Research Methods*, Vol. 7. 29–44.

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

[184] Markus Schedl, Peter Knees, Brian McFee, Dmitry Bogdanov, Bracha Shapira, Lior Rokach, and Francesco Ricci. 2022. Music Recommendation Systems: Techniques, Use Cases, and Challenges. In *Recommender Systems Handbook*. Springer US, New York, NY, 927–971.

[185] Shalom H. Schwartz. 1992. Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. In *Advances in Experimental Social Psychology*. Vol. 25. Elsevier, 1–65. https://doi.org/10.1016/S0065-2601(08)60281-6

[186] Shalom H Schwartz and Wolfgang Bilsky. 1987. Toward A Universal Psychological Structure of Human Values. *Journal of personality and social psychology* 53, 3 (1987), 550–562.

[187] Shalom H Schwartz and Wolfgang Bilsky. 1990. Toward a Theory of the Universal Content and Structure of Values: Extensions and Cross-Cultural Replications. *Journal of personality and social psychology* 58, 5 (1990), 878–891.

[188] Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, and others. 2012. Refining the theory of basic individual values. *Journal of personality and social psychology* 103, 4 (2012), 663.

[189] Shalom H Schwartz and Sipke Huismans. 1995. Value priorities and religiosity in four Western religions. *Social Psychology Quterly* (1995), 88–107.

[190] Yue Shi, Xiaoxue Zhao, Jun Wang, Martha Larson, and Alan Hanjalic. 2012. Adaptive diversification of recommendation results via latent factor portfolio. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, Portland Oregon USA, 175–184. https://doi.org/10.1145/2348283.2348310

[191] Thiago Silveira, Min Zhang, Xiao Lin, Yiqun Liu, and Shaoping Ma. 2019. How good your recommender system is? A survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics* 10, 5 (May 2019), 813–831. https://doi.org/10.1007/s13042-017-0762-9

[192] Nasim Sonboli, Jessie J. Smith, Florencia Cabral Berenfus, Robin Burke, and Casey Fiesler. 2021. Fairness and Transparency in Recommendation: The Users' Perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, Utrecht Netherlands, 274–279. https://doi.org/10.1145/3450613.3456835

[193] Nakarin Sritrakool and Saranya Maneeroj. 2021. Personalized Preference Drift Aware Sequential Recommender System. *IEEE Access* 9 (2021), 155491–155506. https://doi.org/10.1109/ACCESS.2021.3128769

[194] Nakarin Sritrakool, Saranya Maneeroj, and Atsuhiro Takasu. 2025. QUADEN: Discovering Latent Neighbors for Sparse Users and Items across Interaction Quadrants in Recommender System. *ACM Transactions on Information Systems* 43, 4 (July 2025), 1–33. https://doi.org/10.1145/3725886

[195] Abhishek Srivastava, Pradip Kumar Bala, and Bipul Kumar. 2020. New perspectives on gray sheep behavior in E-commerce recommendations. *Journal of Retailing and Consumer Services* 53 (March 2020), 101764. https://doi.org/10.1016/j.jretconser.2019.02.018

[196] Harald Steck. 2010. Training and testing of recommender systems on data missing not at random. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 713–722.

[197] Harald Steck. 2013. Evaluation of recommendations: rating-prediction and ranking. In *Proceedings of the 7th ACM conference on Recommender systems*. ACM, Hong Kong China, 213–220. https://doi.org/10.1145/2507157.2507160

[198] David J Stillwell and Michal Kosinski. 2004. myPersonality project: Example of successful utilization of online social networks for large-scale social research. *American Psychologist* 59, 2 (2004), 93–104.

[199] Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Chloe Bakalar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, Connie Moon Sehat, Sara Johansen, Lianne Kerlin, David Vickrey, Spandana Singh, Sanne Vrijenhoek, Amy Zhang, McKane Andrus, Natali Helberger, Polina Proutskova, Tanushree Mitra, and Nina Vasan. 2024. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *ACM Transactions on Recommender Systems* 2, 3 (Sept. 2024), 1–57. https://doi.org/10.1145/3632297

[200] Mariarosaria Taddeo and Luciano Floridi. 2018. How AI can be a force for good. *Science* 361, 6404 (Aug. 2018), 751–752. https://doi.org/10.1126/science.aat5991

[201] Eugene Tartakovsky. 2023. The psychology of romantic relationships: motivations and mate preferences. *Frontiers in psychology* 14 (2023), 1273607.

[202] Igor V. Tetko, Ruud Van Deursen, and Guillaume Godin. 2024. Be aware of overfitting by hyperparameter optimization! *Journal of Cheminformatics* 16, 1 (Dec. 2024), 139. https://doi.org/10.1186/s13321-024-00934-w

[203] Richard H. Thaler and Cass R. Sunstein. 2021. *Nudge: the Final Edition* (updated edition. ed.). Penguin Books, an imprint of Penguin Random House LLC, New York. Publication Title: Nudge : the final edition.

[204] Mike Thelwall. 2019. Reader and author gender and genre in Goodreads. *Journal of Librarianship and Information Science* 51, 2 (June 2019), 403–430. https://doi.org/10.1177/0961000617709061

[205] Mike Thelwall and Kayvan Kousha. 2017. Goodreads: A social network site for book readers. *Journal of the Association for Information Science and Technology* 68, 4 (April 2017), 972–983. https://doi.org/10.1002/asi.23733

[206] Nava Tintarev, Matt Dennis, and Judith Masthoff. 2013. Adapting Recommendation Diversity to Openness to Experience: A Study of Human Behaviour. In *User Modeling, Adaptation, and Personalization*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Sandra Carberry, Stephan Weibelzahl, Alessandro Micarelli, and Giovanni Semeraro (Eds.). Vol. 7899. Springer Berlin Heidelberg, Berlin, Heidelberg, 190–202. https://doi.org/10.1007/978-3-642-38844-6_16 Series Title: Lecture Notes in Computer Science.

[207] Marko Tkalčič, Li Chen, Bracha Shapira, Lior Rokach, and Francesco Ricci. 2022. Personality and Recommender Systems. In *Recommender Systems Handbook*. Springer US, New York, NY, 757–787.

[208] Yu Tokutake, Kazushi Okamoto, Kei Harada, Atsushi Shibata, and Koki Karube. 2025. A Universal Framework for Offline Serendipity Evaluation in Recommender Systems via Large Language

Models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. ACM, Seoul Republic of Korea, 5294–5298. https://doi.org/10.1145/3746252.3760911

[209] Yoji Tomita, Riku Togashi, and Daisuke Moriwaki. 2022. Matching Theory-based Recommender Systems in Online Dating. In *Proceedings of the 16th ACM Conference on Recommender Systems*. ACM, Seattle WA USA, 538–541. https://doi.org/10.1145/3523227.3547406

[210] Raül Tormos, Christin-Melanie Vauclair, and Henrik Dobewall. 2017. Does Contextual Change Affect Basic Human Values? A Dynamic Comparative Multilevel Analysis Across 32 European Countries. *Journal of Cross-Cultural Psychology* 48, 4 (May 2017), 490–510. https://doi.org/10.1177/0022022117692675

[211] Lourdes Torres. 2004. Accounting and accountability: recent developments in government financial information systems. *Public Administration and Development* 24, 5 (2004), 447–456. https://doi.org/10.1002/pad.332 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pad.332.

[212] Tin T. Tran, Loc Tan Nguyen, and Nguyen Ngoc Phien. 2026. Graph Neural Networks for Collaborative Filtering: A Survey on Ranking Prediction. *IEEE Access* 14 (2026), 11953–11998. https://doi.org/10.1109/ACCESS.2026.3656421

[213] Samira Vaez Barenji, Sushobhan Parajuli, and Michael D. Ekstrand. 2025. User and Recommender Behavior Over Time: Contextualizing Activity Effectiveness Diversity and Fairness in Book Recommendation. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. ACM, New York City USA, 280–287. https://doi.org/10.1145/3708319.3733710

[214] Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2018. On the robustness and discriminative power of information retrieval metrics for top-N recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, Vancouver British Columbia Canada, 260–268. https://doi.org/10.1145/3240323.3240347

[215] Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2020. Assessing ranking metrics in top-N recommendation. *Information Retrieval Journal* 23, 4 (Aug. 2020), 411–448. https://doi.org/10.1007/s10791-020-09377-x

[216] Alejandro Valencia-Arias, Hernán Uribe-Bedoya, Juan David González-Ruiz, Gustavo Sánchez Santos, Edgard Chapoñan Ramírez, and Ezequiel Martínez Rojas. 2024. Artificial intelligence and recommender systems in e-commerce. Trends and research agenda. *Intelligent Systems with Applications* 24 (Dec. 2024), 200435. https://doi.org/10.1016/j.iswa.2024.200435

[217] Hester Van Herk, Ype H. Poortinga, and Theo M. M. Verhallen. 2004. Response Styles in Rating Scales: Evidence of Method Bias in Data From Six EU Countries. *Journal of Cross-Cultural Psychology* 35, 3 (May 2004), 346–360. https://doi.org/10.1177/0022022104264126

[218] Saúl Vargas, Linas Baltrunas, Alexandros Karatzoglou, and Pablo Castells. 2014. Coverage, redundancy and size-awareness in genre diversity for recommender systems. In *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, Foster City, Silicon Valley California USA, 209–216. https://doi.org/10.1145/2645710.2645743

[219] Saúl Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, Chicago Illinois USA, 109–116. https://doi.org/10.1145/2043932.2043955

[220] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *Advances in neural information processing systems* 30 (2017).

[221] Michele Vecchione. 2023. The five factors of personality and personal values: An update with the refined theory. *Personality and Individual Differences* 203 (2023), 112033.

[222] Katrien Verbert, Erik Duval, Stefanie N Lindstaedt, and Denis Gillet. 2010. Context-aware recommender systems. *Journal of Universal Computer Science* 16, 16 (2010), 2175–2178.

[223] Bas Verplanken and Rob W. Holland. 2002. Motivated decision making: Effects of activation and self-centrality of values on choices and behavior. *Journal of Personality and Social Psychology* 82, 3 (2002), 434–447. https://doi.org/10.1037/0022-3514.82.3.434

[224] E.M. Voorhees. 1999. The TREC-8 question answering track report. In *Proc. TREC*. National Institute of Standards and Technology, 77–82.

[225] RJJ Voorn, G Van der Veen, TJL Van Rompay, SM Hegner, and Aadriaan TH Pruyn. 2021. Human values as added value(s) in consumer brand congruence: a comparison with traits and functional requirements. *Journal of Brand management* 28, 1 (2021), 48–59.

[226] Mengting Wan and Julian J. McAuley. 2018. 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*, Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan (Eds.). ACM, 86–94. https://doi.org/10.1145/3240323.3240369

[227] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. 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*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 2605–2610. https://doi.org/10.18653/V1/P19-1248

[228] Jianling Wang, Kaize Ding, Liangjie Hong, Huan Liu, and James Caverlee. 2020. Next-item Recommendation with Sequential Hypergraphs. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Virtual Event China, 1101–1110. https://doi.org/10.1145/3397271.3401133

[229] Jianling Wang, Ziwei Zhu, and James Caverlee. 2020. User Recommendation in Content Curation Platforms. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. ACM, Houston TX USA, 627–635. https://doi.org/10.1145/3336191.3371822

[230] Shoujin Wang, Longbing Cao, Yan Wang, Quan Z. Sheng, Mehmet A. Orgun, and Defu Lian. 2022. A Survey on Session-based Recommender Systems. *Comput. Surveys* 54, 7 (Sept. 2022), 1–38. https://doi.org/10.1145/3465401

[231] Siyu Wang, Xiaocong Chen, Dietmar Jannach, and Lina Yao. 2023. Causal Decision Transformer for Recommender Systems via Offline Reinforcement Learning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Taipei Taiwan, 1599–1608. https://doi.org/10.1145/3539618.3591648

[232] Siyu Wang, Xiaocong Chen, and Lina Yao. 2025. Retentive Decision Transformer with Adaptive Masking for Reinforcement Learning-Based Recommendation Systems. *ACM Transactions on Intelligent Systems and Technology* 16, 3 (June 2025), 1–20. https://doi.org/10.1145/3719208

[233] Shanfeng Wang, Maoguo Gong, Haoliang Li, and Junwei Yang. 2016. Multi-objective optimization for long tail recommendation. *Knowledge-Based Systems* 104 (July 2016), 145–155. https://doi.org/10.1016/j.knosys.2016.04.018

[234] Shanfeng Wang, Maoguo Gong, Yue Wu, and Mingyang Zhang. 2020. Multi-objective optimization for location-based and preferences-aware recommendation. *Information Sciences* 513 (March 2020), 614–626. https://doi.org/10.1016/j.ins.2019.11.028

[235] Di Wu, Xin Luo, Mingsheng Shang, Yi He, Guoyin Wang, and MengChu Zhou. 2021. A Deep Latent Factor Model for High-Dimensional and Sparse Matrices in Recommender Systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 51, 7 (July 2021), 4285–4296. https://doi.org/10.1109/TSMC.2019.2931393

[236] Wen Wu, Li Chen, and Yu Zhao. 2018. Personalizing recommendation diversity based on user personality. *User Modeling and User-Adapted Interaction* 28, 3 (Aug. 2018), 237–276. https://doi.org/10.1007/s11257-018-9205-x

[237] Ruobing Xie, Yanlei Liu, Shaoliang Zhang, Rui Wang, Feng Xia, and Leyu Lin. 2021. Personalized Approximate Pareto-Efficient Recommendation. In *Proceedings of the Web Conference 2021*. ACM, Ljubljana Slovenia, 3839–3849. https://doi.org/10.1145/3442381.3450039

[238] Yueqi Xie, Peilin Zhou, and Sunghun Kim. 2022. Decoupled Side Information Fusion for Sequential Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Madrid Spain, 1611–1621. https://doi.org/10.1145/3477495.3531963

[239] Xin Luo, Mengchu Zhou, Yunni Xia, and Qingsheng Zhu. 2014. An Efficient Non-Negative Matrix-Factorization-Based Approach to Collaborative Filtering for Recommender Systems. *IEEE Transactions on Industrial Informatics* 10, 2 (May 2014), 1273–1284. https://doi.org/10.1109/TII.2014.2308433

[240] Lanling Xu, Zhen Tian, Gaowei Zhang, Lei Wang, Junjie Zhang, Bowen Zheng, Yifan Li, Yupeng Hou, Xingyu Pan, Yushuo Chen, Wayne Xin Zhao, Xu Chen, and Ji-Rong Wen. 2022. Recent Advances in RecBole: Extensions with more Practical Considerations.

[241] Emre Yalcin and Alper Bilge. 2022. Evaluating unfairness of popularity bias in recommender systems: A comprehensive user-centric analysis. *Information Processing & Management* 59, 6 (Nov. 2022), 103100. https://doi.org/10.1016/j.ipm.2022.103100

[242] Eva Zangerle and Christine Bauer. 2023. Evaluating Recommender Systems: Survey and Framework. *Comput. Surveys* 55, 8 (Aug. 2023), 1–38. https://doi.org/10.1145/3556536

[243] Kui Zhang, Jianguang Hu, and Xing Xin. 2023. Deep Interest Network Based Book Recommends-A Case Study of College Reader. In *2023 8th International Conference on Information Systems Engineering*. ACM, Bangkok Thailand, 84–89. https://doi.org/10.1145/3641032.3641061

[244] Shuai Zhang, Yi Tay, Lina Yao, Aixin Sun, Ce Zhang, Bracha Shapira, Lior Rokach, and Francesco Ricci. 2022. Deep Learning for Recommender Systems. In *Recommender Systems Handbook*. Springer US, New York, NY, 173–210.

[245] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2020. Deep Learning Based Recommender System: A Survey and New Perspectives. *Comput. Surveys* 52, 1 (Jan. 2020), 1–38. https://doi.org/10.1145/3285029

[246] Yuan Cao Zhang, Diarmuid O Seaghdha, Daniele Quercia, and Tamas Jambor. 2012. Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, Seattle Washington USA, 13–22. https://doi.org/10.1145/2124295.2124300

[247] Qian Zhao, F. Maxwell Harper, Gediminas Adomavicius, and Joseph A. Konstan. 2018. Explicit or implicit feedback? engagement or satisfaction?: a field experiment on machine-learning-based recommender systems. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. ACM, Pau France, 1331–1340. https://doi.org/10.1145/3167132.3167275

[248] Wayne Xin Zhao, Yupeng Hou, Xingyu Pan, Chen Yang, Zeyu Zhang, Zihan Lin, Jingsen Zhang, Shuqing Bian, Jiakai Tang, Wenqi Sun, et al. 2022. RecBole 2.0: Towards a More Up-to-Date Recommendation Library. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4722–4726.

[249] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. 2021. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms. In *CIKM*. ACM, 4653–4664.

[250] Yong Zheng, Archana Subramaniyan, Leonard Barolli, Fatos Xhafa, Makoto Takizawa, and Tomoya Enokido. 2019. Personality-Aware Collaborative Learning: Models and Explanations. In *Advanced Information Networking and Applications*. Advances in Intelligent Systems and Computing, Vol. 926. Springer International Publishing AG, Switzerland, 631–642.

[251] Yanbo Zhou, Gang-Feng Ma, Xilin Wen, Xu-Hua Yang, and Yi-Cheng Zhang. 2026. Sequential recommender systems: A methodological taxonomy and research frontiers. *Computer Science Review* 59 (Feb. 2026), 100818. https://doi.org/10.1016/j.cosrev.2025.100818

[252] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web - WWW '05*. ACM Press, Chiba, Japan, 22. https://doi.org/10.1145/1060745.1060754

[253] Zainab Zolaktaf, Reza Babanezhad, and Rachel Pottinger. 2018. A Generic Top-N Recommendation Framework for Trading-Off Accuracy, Novelty, and Coverage. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE, Paris, 149–160. https://doi.org/10.1109/ICDE.2018.00023

[254] Feng Zou, Debao Chen, Qingzheng Xu, Ziqi Jiang, and Jiahui Kang. 2021. A two-stage personalized recommendation based on multi-objective teaching–learning-based optimization with decomposition. *Neurocomputing* 452 (Sept. 2021), 716–727. https://doi.org/10.1016/j.neucom.2020.08.080

[255] Erion Çano and Maurizio Morisio. 2017. Hybrid recommender systems: A systematic literature review. *Intelligent data analysis* 21, 6 (2017), 1487–1524.

A

Value Profile Construction Examples

A.1. Definition of Schwartz's basic human values

For the user value modeling, we used the theory of basic human values, developed by Shalom H. Schwartz [185]. This cross-cultural theory identified 10 basic human values, which can be organized in five higher-order groups.[1] These values, grouped by their higher-order category, are defined below. All definitions are derived from Schwartz [185].

A.1.1. Openness to change

- Self-direction: “The defining goal of this value type is independent thought and action - choosing, creating, exploring”
- Stimulation: “Stimulation values derive from the presumed organismic need for variety and stimulation in order to maintain an optimal level of activation”

A.1.2. Conservation

- Conformity: “The defining goal of this value type is restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms.”
- Security: “The motivational goal of this value type is safety, harmony, and stability of society, of relationships, and of self.”
- Tradition: “The motivational goal of tradition values is respect, commitment, and acceptance of the customs and ideas that one’s culture or religion impose on the individual (respect for tradition, humble, devout, accepting my portion in life, moderate).”

A.1.3. Self-enhancement

- Achievement: “The defining goal of this value type is personal success through demonstrating competence according to social standards”
- Power: “We view the central goal of power values as attainment of social status and prestige, and control or dominance over people and resources (authority, wealth, social power, preserving my public image, social recognition).”

A.1.4. Self-transcendence

- Benevolence: “The motivational goal of benevolence values is preservation and enhancement of the welfare of people with whom one is in frequent personal contact (helpful, loyal, forgiving, honest, responsible, true friendship, mature love).”
- Universalism: “The motivational goal of universalism is understanding, appreciation, tolerance, and protection for the welfare of all people and for nature.”

[1] Note that Hedonism is sometimes included in the higher-order group of self-enhancement, but according to the original paper, it should be its own group.

A.1.5. Hedonism

Hedonism: “we can define the motivational goal of this type more sharply as pleasure or sensuous gratification for oneself (pleasure, enjoying life)” [185]

A.2. Text Value Profile Example (Review/Book Description)

We will use the following review[2] to illustrate the construction of a Value Profile for a Text:

> Jack’s budding (god, i hate that i just used that word) appreciation for poetry, and confidence in his own writing grows at a believable pace, and his excitement about what he reads and his shyness concerning his own words are both achingly sweet (and familiar) sentiments. Relatable read for kids, sweet/funny read for adults, like reading a journal you wrote in elementary school; a little embarassing but at moments striking in its simplicity (in this case: When Jack delikes a poem because he doesn’t understand it, then decides that maybe it doesn’t have to make sense. Maybe the author was just trying to make a picture with words. Simple. Perfect. Made me say “Oh, right. Exactly.” to myself.
>
> jacks budding god i hate that i just used that word appreciation for poetry and confidence in his own writing grows at a believable pace and his excitement about what he reads and his shyness concerning his own words are both achingly sweet and familiar sentiments relatable read for kids sweetfunny read for adults like reading a journal you wrote in elementary school a little embarassing but at moments striking in its simplicity in this case when jack delikes a poem because he doesnt understand it then decides that maybe it doesnt have to make sense maybe the author was just trying to make a picture with words simple perfect made me say oh right exactly to myself

The value words are color coded in the review below, and Table A.1 gives the counts for each value.

> jacks budding god i hate that i just used that word appreciation for poetry and confidence in his own writing grows at a believable pace and his excitement about what he reads and his shyness concerning his own words are both achingly sweet and familiar sentiments relatable read for kids sweetfunny read for adults like reading a journal you wrote in elementary school a little embarassing but at moments striking in its simplicity in this case when jack delikes a poem because he doesnt understand it then decides that maybe it doesnt have to make sense maybe the author was just trying to make a picture with words simple perfect made me say oh right exactly to myself

Table A.1: Count of value words in the review

Basic Human Value	Words in Text	Count
Security	-	0
Conformity	’shyness’, ’familiar’	2
Tradition	’god’	1
Benevolence	’confidence’	1
Universalism	’understand’	1
Self-Direction	-	0
Stimulation	’excitement’	1
Hedonism	-	0
Achievement	’appreciation’	1
Power	-	0

The total number of value-related words is calculated as the sum of the counts, which is 7 in this case. Therefore, we divide all the counts by 7, which gives us the normalized count shown in Table A.2.

[2] review_id $7f8034f2b7218a8b8ce1b48997db2c3d$

Table A.2: Normalized count (frequency) of value words in the review

Basic Human Value	Normalized Count
Security	0
Conformity	2/7
Tradition	1/7
Benevolence	1/7
Universalism	1/7
Self-Direction	0
Stimulation	1/7
Hedonism	0
Achievement	1/7
Power	0

This matches the paper's description: "The final value score is calculated as the frequency of the words representing the given value, minus the frequency of all value-related words in the text." [169]. We calculate the mean frequency with Equation A.1, and subtract this from the normalized counts from Table A.2.

$$\text{Mean frequency} = \frac{0 + \frac{2}{7} + \frac{1}{7} + \frac{1}{7} + \frac{1}{7} + 0 + \frac{1}{7} + 0 + \frac{1}{7} + 0}{10} = \frac{1}{10} = 0.1 \tag{A.1}$$

Table A.3 shows the resulting Review Value Profile of the review. To better understand the differences in frequency between the values, we visualize this Review Value Profile as a bar plot in Figure A.1. We can see that words related to the value Conformity are more frequent than other value-related words in the text.

Table A.3: Review Value Profile for review

Value	Ipsatized score
Security	-0.1
Conformity	0.185
Tradition	0.042
Benevolence	0.042
Universalism	0.042
Self-direction	-0.1
Stimulation	0.042
Hedonism	-0.1
Achievement	0.042
Power	-0.1

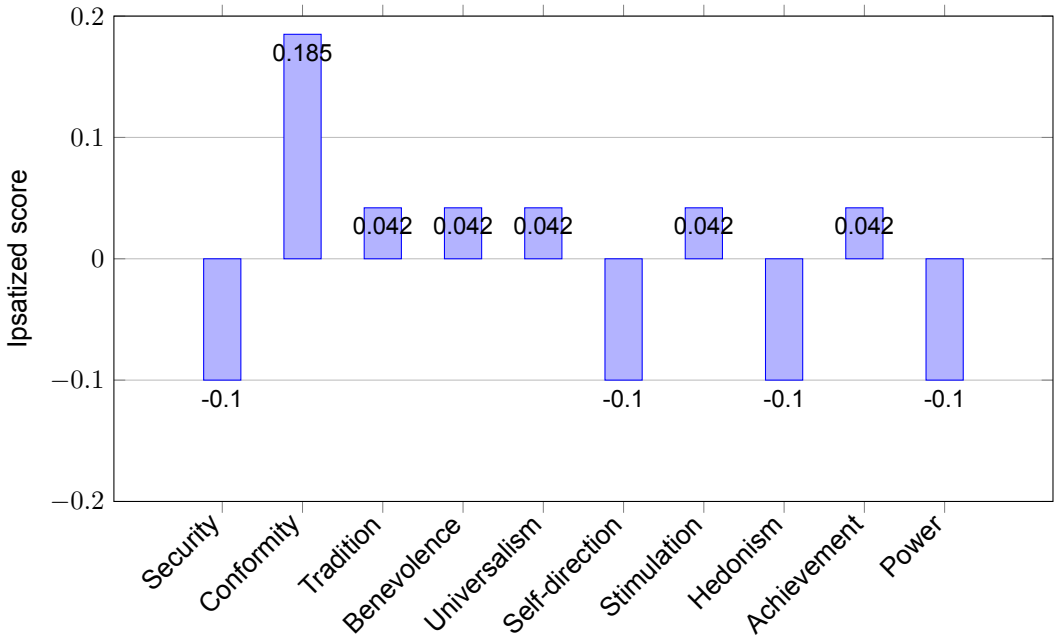


Figure A.1: Bar plot of the Review Value Profile for review

A.3. User/Recommendation List Value Profile Example

Building on the example given in section 5.2, we focus on the user that the review[3] discussed above belongs to[4]. The User Value Profile for this user can be seen in Table A.4, with a bar plot representation in Figure A.2.

Table A.4: Value profile for example user from the Goodreads Poetry dataset

Value	Ipsatized score
Security	-0.079
Conformity	-0.031
Tradition	-0.076
Benevolence	0.250
Universalism	-0.020
Self-direction	-0.037
Stimulation	-0.020
Hedonism	0.066
Achievement	0.048
Power	-0.099

[3] review_id $28423ff309bc896c071a8d9df4a10e8a$
[4] user with user_id $008ffafc7ea81f88131f5a254a8cef89$

Figure A.2: Bar plot of the Review Value Profile for review

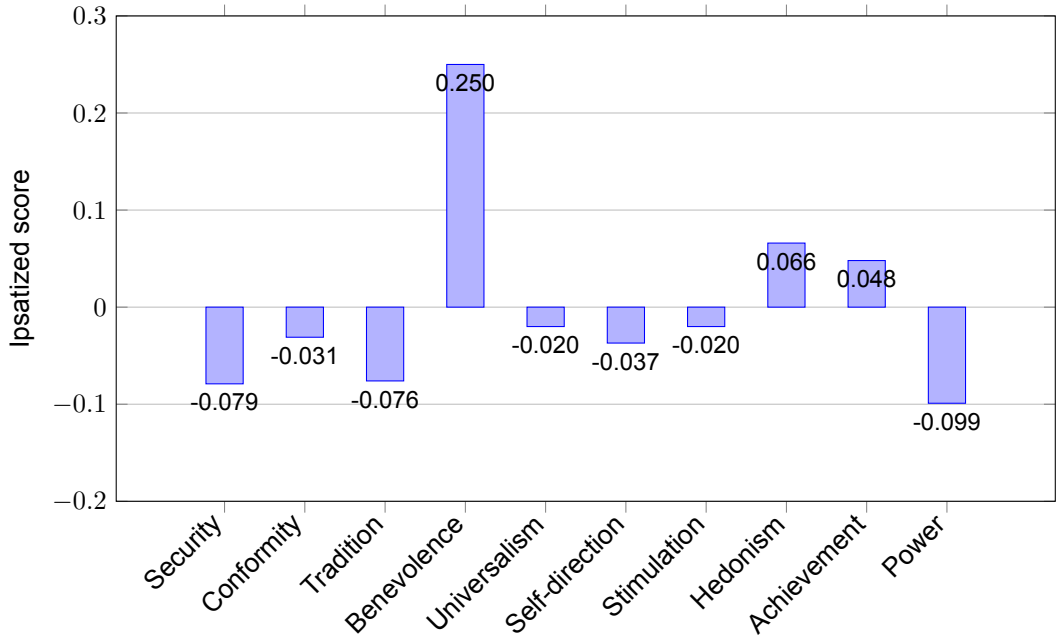
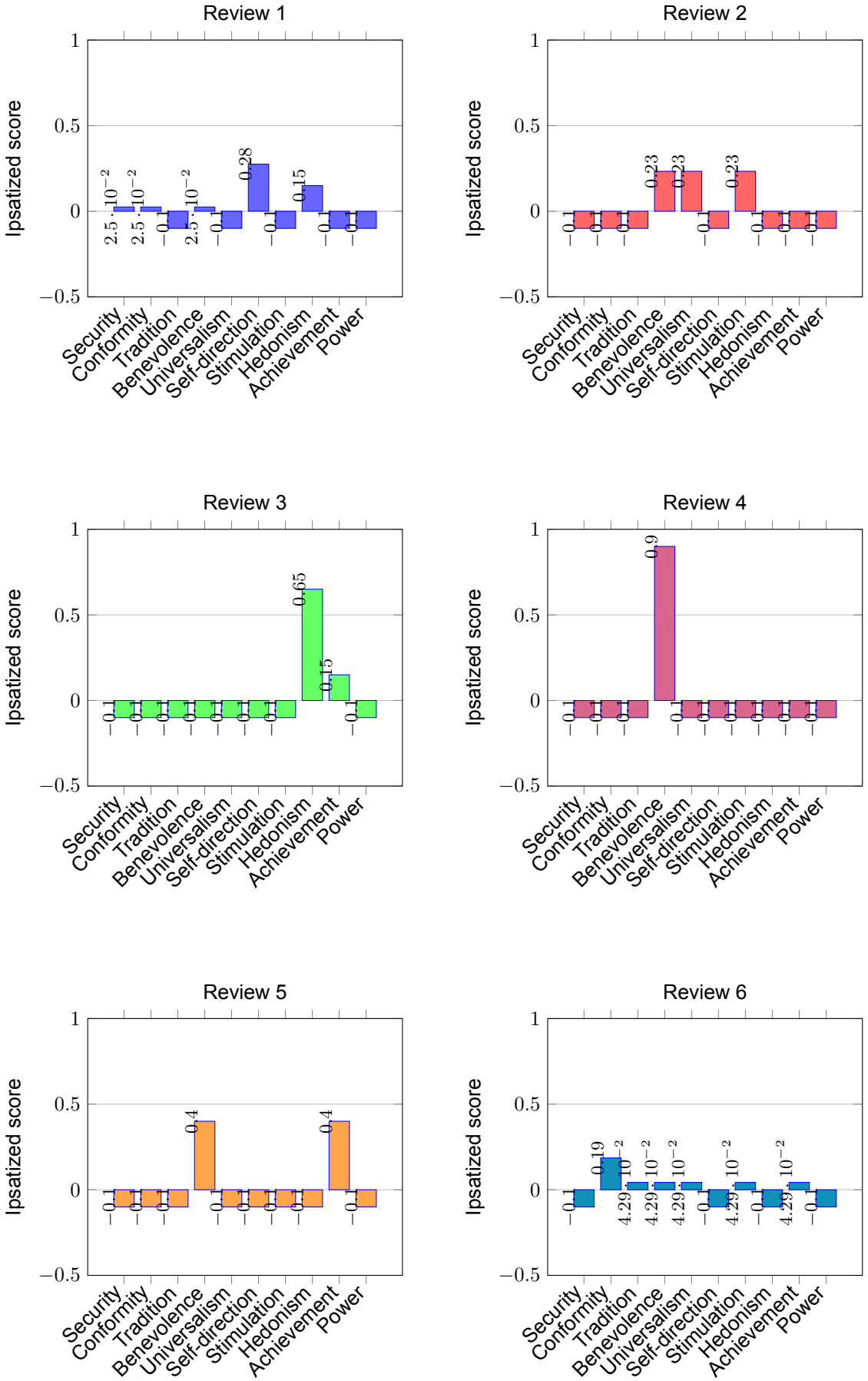


Figure A.3: Review value profiles for a user from the Goodreads Poetry dataset



The resulting distributions of ipsatization reflect relative value preferences indicated in that review, not absolute levels. This ipsatization is necessarily to reflect the theoretical basis of the theory of basic human values, namely that values are ordered by importance relative to each other. Without ipsatization, raw scores might reflect a tendency to use more value-related words overall. Ipsatization removes this bias, focusing on the pattern of values rather than their absolute frequency. Thus, the values with a higher positive score are predicted to be more important to a user.

B

Configuration files

B.1. Finding the number of epochs experiment (fm_goodreads_uvp.yaml

```
SER_ID_FIELD: user_id
ITEM_ID_FIELD: item_id
RATING_FIELD: rating

load_col:
  inter: [user_id, item_id, rating]
  user: [user_id, Security, Conformity, Tradition, Benevolence, Universalism, Self-Direction, Stimul

numerical_features: ['Security', 'Conformity', 'Tradition', 'Benevolence', 'Universalism', 'Self-Dir

# Filter users who have at least 4 interactions (based on exploratory experiment)
user_inter_num_interval: "[4,inf)"

# ======================
# General
# ======================
use_gpu: True
distributed: True
worker: 24
seed: 2020
reproducibility: True
state: INFO
show_progress: True
save_model: False
load_best_model: False

save_dataset: True
enable_amp: True
enable_scaler: True

# ======================
# Training
# ======================
epochs: 10
train_batch_size: 262144
eval_batch_size: 262144
```

```
embedding_size: 32

train_neg_sample_args:
  distribution: uniform
  sample_num: 1
  alpha: 1.0
  dynamic: False
  candidate_num: 0

eval_step: 1
stopping_step: 5
learning_rate: 0.02 # Recommended for FM

# ======================
# Evaluation
# ======================

eval_args:
  split: {'RS': [8,1,1]}
  group_by: user
  order: RO
  mode:
    valid: uni100
    test: uni100

metrics: ['NDCG']
topk: [10]
valid_metric: NDCG@10
metric_decimal_place: 4
```

C

Investigated Hyperparameters

C.1. UserKNN

```
k choice [10,50,100,200,250,300,400,500,1000,1500,2000,2500]
shrink choice [0.0,1.0]
```

C.2. BPRMF

```
learning_rate choice [0.01,0.005,0.001,0.0005,0.0001]
```

C.3. DeepFM

```
learning_rate choice [0.01,0.005,0.001,0.0005,0.0001]
dropout_prob choice [0.0,0.1,0.2,0.3,0.4,0.5]
mlp_hidden_size choice ['[64,64,64]','[128,128,128]','[256,256,256]','[512,512,512]']
```

C.4. SASRec

```
learning_rate choice [0.01,0.005,0.001,0.0005,0.0001]
attn_dropout_prob choice [0.2, 0.5]
hidden_dropout_prob choice [0.2, 0.5]
n_heads choice [1, 2]
n_layers choice [1,2,3]
```