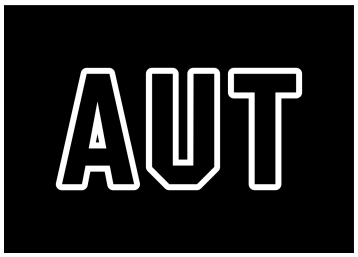


Advancing Responsible Recommendation Systems: Enhancing Accuracy, Diversity, and Fairness

Mengyan Wang

Supervisor: Dr Weihua Li
A/Prof. Quan Bai
A/Prof. Jian Yu

School of Engineering, Computer & Mathematical Sciences
Auckland University of Technology

A thesis submitted to Auckland University of Technology
in fulfilment of the requirements for the degree of
Doctor of Philosophy

May 2025

Dedication

To my dearest mother, Chunrong Yan, and my partner, Shun Lyu, thank you for your continued support and encouragement over the years. Your love and understanding have laid the foundation for my academic journey. During challenging times, your presence gave me the strength to persevere. I could not have completed this journey without your selfless dedication and love. I dedicate this work to you both, with deep gratitude for all you have given me.

Copyright

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgments.

Mengyan Wang
May 2025

Co-authorship Contribution

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgments.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Dr. Weihua Li, A/Prof. Quan Bai, and A/Prof. Jian Yu, for their invaluable guidance, support, and encouragement throughout my Ph.D. journey. Their expertise, dedication, and patience played an instrumental role in the successful completion of my research. I am profoundly grateful for the time and effort they have invested in me, and for sharing their immense knowledge and skills. Their insightful feedback and constructive criticism have greatly enhanced my research ability. I especially appreciate their support during the challenging times, which was crucial to achieving my academic aspirations.

I would also like to extend my heartfelt appreciation to Ms. Christine Bear for her continuous encouragement. Her unwavering support not only contributed to my research but also instilled in me the confidence and motivation to face the challenges of this journey. Christine's belief in my capabilities has been a significant driving force behind both my academic and professional growth.

Furthermore, I acknowledge the financial support provided by Callaghan Innovation (CSITR1901, 2021), New Zealand, without which this research would not have been possible. I am deeply appreciative of their contributions to advancing science and technology in New Zealand. I would also like to thank CAITO.AI for their invaluable partnership and contributions to the research.

I am equally grateful to my colleagues, Dr. Shiqing Wu, Dr. Yuxuan Hu, Dr. Jingli Shi, and Guan Wang, for their support throughout my Ph.D. journey. Their friendship has been a constant source of motivation and inspiration, and their contributions to my research have been of immense value.

Abstract

With the rapid growth of personalized content delivery, recommendation systems have become integral to shaping user experiences across various platforms. However, despite their widespread use, existing RSs face persistent challenges related to accuracy, diversity, and fairness. These issues raise ethical concerns and compromise the overall performance of RSs, affecting user trust and satisfaction.

Traditional recommendation systems, primarily based on collaborative filtering, content-based approaches, and more recent artificial intelligence-driven models, have indeed personalized content delivery by providing recommendations tailored to individual user preferences. However, these systems continue to deal with continuous challenges such as the cold start problem, data sparsity, over-specialization, and lack of fairness. These limitations compromise the quality of recommendations and weaken user trust and satisfaction, potentially leading to reduced user engagement. Facing these challenges, this thesis introduces three novel responsible models to enhance recommendation systems' performance and ethical standards.

The Dual Observation-Based Recommendation (DOR) model is first introduced, integrating local and global observation mechanisms to enhance recommendation accuracy and diversity. This approach is particularly effective in addressing challenges such as data sparsity, cold starts, and filter bubbles. By utilizing a broader range of contextual data, the DOR model gains a deeper insight into user preferences, resulting in more precise and personalized recommendations. Additionally, the inclusion of external information helps to alleviate filter bubble issues.

Compared to the DOR model, the Responsible Graph-Based Recommendation (RGRec) model emphasizes addressing the issue of filter bubbles. RGRec utilizes innovative strategies, such as belief nudging and generative AI, to ensure that users are exposed to a broader range of content, promoting engagement with diverse perspectives. This design aims to reduce the risks associated with reinforcing existing biases and to foster belief harmony among users. By enhancing content exposure, RGRec effectively mitigates the adverse effects of filter bubbles.

Third, inspired by Yin-Yang theory, the Agent-Based Adaptive Information Neutrality (AAIN) model introduces a multi-agent framework that dynamically adjusts information exposure to mitigate recommendation biases and ensure a neutral, diverse recommendation environment. The proposed AAIN model adapts its recommendations by balancing different sentiment information and enhancing content diversity without compromising accuracy.

Extensive experimental evaluations demonstrate that these models significantly improve RS performance regarding accuracy, diversity, and fairness. They effectively address the cold start problem, data sparsity, filter bubbles, and lack of fairness, which is critical for fostering user trust.

This thesis advances the technical capabilities of recommendation systems and highlights the importance of incorporating ethical considerations into their design. By contributing to the broader discourse on responsible artificial intelligence, this thesis emphasizes the need for developing recommendation systems that prioritize ethical outcomes alongside technical performance. The development of the DOR, RGRec, and AAIN models lays a strong foundation for future recommendation systems, ensuring they evolve to enhance user trust and satisfaction while remaining technologically advanced and ethically responsible.

Publications

M. Wang, W. Li, J. Shi, S. Wu, and Q. Bai, "DOR: A novel dual-observation-based approach for recommendation systems," Applied Intelligence, vol. 53, no. 23, pp. 29109-29127, 2023.

M. Wang, Y. Hu, S. Wu, W. Li, Q. Bai, Z. Yuan, and C. Jiang, "Nudging towards responsible recommendations: A graph-based approach to mitigate belief filter bubbles," IEEE Transactions on Artificial Intelligence, vol. 6, no. 2, pp. 378–392, Feb. 2025, doi: 10.1109/TAI.2024.3373392.

M. Wang, Y. Hu, S. Wu, W. Li, Q. Bai, and V. Rupar, "Balancing information perception with Yin-Yang: Agent-based information neutrality model for recommendation systems," IEEE Transactions on Computational Social Systems, 2025.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Recommendation Systems (RSs) have become essential in digital platforms, offering users personalized content and aiding decision-making in environments overwhelmed with information. Over time, RSs have evolved from traditional algorithms, such as collaborative filtering (CF) and content-based filtering (CBF), to more sophisticated techniques that take advantage of machine learning (ML) and deep neural networks (DNN) to improve prediction accuracy and system scalability [76, 160].

Despite these advancements, RSs continue to face several foundational challenges that limit their effectiveness and raise broader concerns about their societal impact. These include the cold start problem, data sparsity, lack of diversity, the emergence of filter bubbles, and algorithmic fairness [23]. Such issues not only reduce the quality and relevance of recommendations but also risk diminishing user trust, engagement, and equitable access to information.

Traditional approaches such as CF, CBF, and Hybrid Filtering (HF) have been widely adopted to personalize content. However, they often fail to address these challenges comprehensively. For example, memory-based CF suffers from data sparsity, while CBF can lead to overfitting and reduced content diversity. While ML- and DNN-based models effectively capture complex patterns, they may also reinforce user biases and intensify filter bubbles and fairness issues.

These limitations highlight the urgent need for a new generation of RSs that balance predictive accuracy with ethical considerations such as diversity, fairness, and user autonomy. This thesis addresses these challenges by introducing three novel models, namely the Dual-Observation-based Recommendation (DOR), the Responsible Graph-based Recommendation (RGRec), and the Agent-based Adaptive Information Neutralization (AAIN). These models are designed to provide scalable and practical solutions for building more responsible, user-centered RSs.

To contextualize these contributions, this chapter first reviews the evolution and classification of RSs, examining the advantages and limitations of traditional and ML-based methods. It then introduces the notion of responsible RSs and outlines how this thesis proposes to bridge the gap between algorithmic performance and ethical recommendation practices. Together, these discussions set the stage for the research questions and contributions presented in the remainder of the thesis.

1.1 Background of Recommendation Systems

On the Internet, users are confronted with an overwhelming array of options, making it essential to employ efficient methods to filter, prioritize, and deliver relevant information. This is crucial for preventing information overload, which can present significant user challenges [172]. RSs address this issue by leveraging algorithms to provide personalized services tailored to individual preferences [95]. Over the past decades, RSs have been extensively researched and have proven effective across various scenarios. Many prominent online platforms, such as Amazon [1], Netflix [2], and YouTube [3], have adopted RSs to enhance user satisfaction and drive sales [182]. Figure 1.1 presents the basic workflow of an RS, where user interactions, such as searching, liking, and watching content, influence the system's selection of items like songs, movies, or articles. These interactions are processed by the RSs, which generate personalized recommendations and continuously refine the suggestions based on user feedback. Ultimately, the primary function of RSs is to assess the relevance of an item and determine whether it is suitable for recommendation [251], which can be expressed as:

$$f : V \times I \rightarrow R, \tag{1.1}$$

where $V = \{v_1, ..., v_n\}$ represents a set of users and $I = \{i_1, \ldots, i_m\}$ denotes a set of items on the recommendation platform. R represents the predicted ratings for items in I by users in V. The final recommendation list can be generated by ranking the items in I based on these predicted ratings.

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

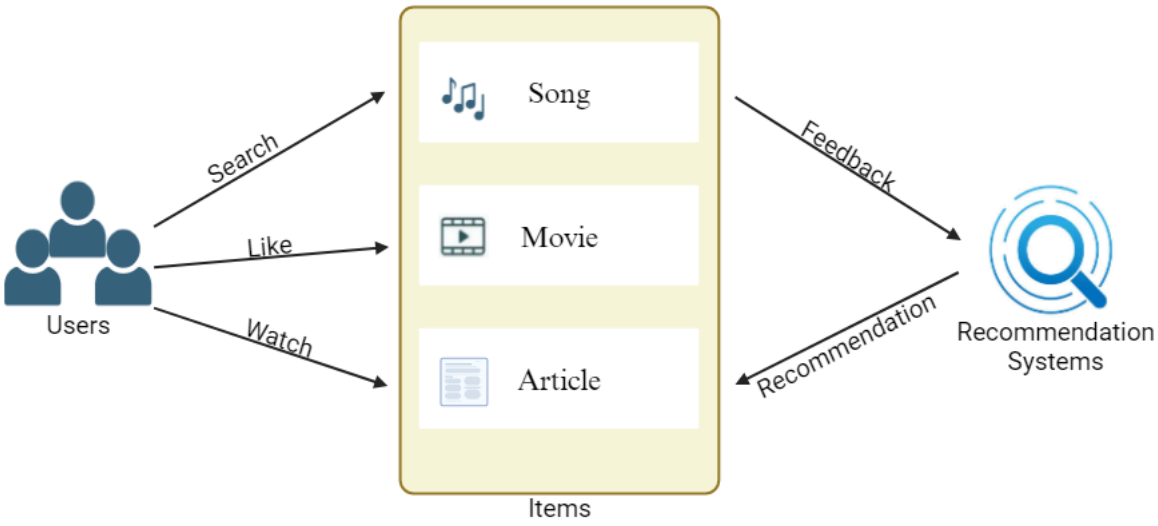


Fig. 1.1: General Recommendation System Working Diagram

1.1.1 Classification of Recommendation Systems

The RSs are typically classified into three main methods, i.e., Collaborative Filtering (CF), Content-Based Filtering (CBF), and Hybrid Filtering (HF) [74], which are demonstrated in Figure 1.2. CF is further divided into Model-based and Memory-based approaches. Traditional CF methods are part of both the Model-based and Memory-based categories. In the Model-based CF branch, there is a distinction between traditional model-based methods, which rely on more straightforward statistical techniques, and more advanced approaches that have evolved, such as Deep Learning (DL). These DL methods represent an evolution from traditional techniques, capable of capturing more complex, non-linear patterns in the data, marking a shift towards more sophisticated and flexible RSs. The figure emphasizes this progression from traditional methods to more advanced, DL-based models. Each method has its strengths and limitations, which will be discussed below.

Collaborative Filtering Method

Collaborative Filtering (CF), a traditional method in RSs, effectively delivers personalized content by analyzing preferences from users with similar historical behaviors or preferences [136]. CF methods can typically be divided into two main types: memory-based and model-based types. Memory-based methods use interaction data to calculate recommendations based on the similarities between users or items. In contrast, model-based methods focus on generating a summarized representation of rating patterns, often done offline [5].

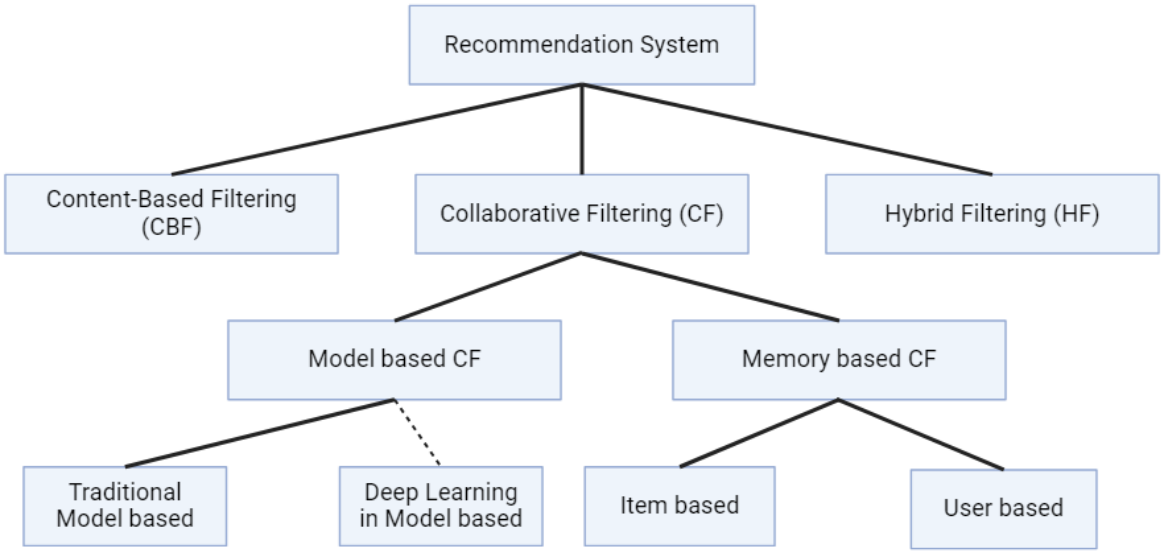


Fig. 1.2: Recommendation Methods Structure

Memory-based CF methods utilize past data to estimate unknown ratings by applying heuristics. These methods, also known as neighborhood-based filtering, are based on the assumption that users with similar historical preferences will continue to display similar preferences in the future. In this method, ratings for items are predicted by aggregating the ratings of similar users or items [141]. A common approach in memory-based methods is the nearest-neighbor technique, which relies on various distance measures [168]. This approach predicts ratings by identifying users with preferences similar to the target user's and leveraging neighborhood-based techniques typical of memory-based CF. To produce predictions, these methods compute similarities across the user-item database, grouping users with similar preferences [203]. K-Nearest-Neighbor (KNN) is a classic example of this approach. KNN is user-friendly and can seamlessly integrate new data without depending on the content of the recommended items, which enhances its ability to provide accurate ratings across a broader range of items. Memory-based methods are easy to implement, making them a practical choice for quickly addressing problems with moderately sized datasets [48].

However, memory-based methods have notable drawbacks. **Sparsity:** As datasets grow more prominent, the sparsity of user-item interactions increases, making it harder to find reliable neighbors, leading to scalability challenges. **Cold Start:** Memory-based systems also struggle with cold start scenarios, where limited interactions with new users or items hinder accurate recommendations. Additionally, **Limited Representation:** These methods often fail to capture complex patterns in the data, such as non-linear relationships, which may result in less effective recommendations [16].

As previously discussed, memory-based CF methods rely on historical data to predict ratings, commonly using nearest-neighbor techniques to group users with similar interests [151]. These methods can be implemented using various approaches, including user-based and item-based CF methods. The item-based CF method consists of two main phases: first, identifying similarities between items using different similarity measures, and second, predicting ratings for unknown items based on these similarity scores. Item-based CF method is conceptually similar to the user-based CF method, but instead of calculating user similarities, it focuses on item similarities Sim_{item} [90]. A commonly used method to calculate item similarity is cosine similarity, as illustrated in Equation 1.2.

$$\text{Sim}_{\text{item}}(i_1, i_2) = \frac{\sum_{v_k \in V} R_{v_k,i_1} \cdot R_{v_k,i_2}}{\sqrt{\sum_{v_k \in V} R^2_{v_k,i_1}} \cdot \sqrt{\sum_{v_k \in V} R^2_{v_k,i_2}}} \tag{1.2}$$

Here, R_{v_k,i_1} and R_{v_k,i_2} represent the ratings given by user v_k to items i_1 and i_2, respectively. The cosine similarity measures the cosine of the angle between the rating vectors of the two items. Additionally, the prediction for item-based CF methods for user v_k and item i_2 is calculated as shown in Equation 1.3.

$$\text{P}_{\text{item-based}}(v_k, i_2) = \frac{\sum_{i_1 \in R_{v_k}} \text{Sim}_{\text{item}}(i_1, i_2) \cdot R_{v_k,i_1}}{\sum_{i_1 \in R_{v_k}} \text{Sim}_{\text{item}}(i_1, i_2)} \tag{1.3}$$

$\text{P}_{\text{item-based}}(v_k, i_2)$ represents the predicted rating for user v_k on item i_2 using the item-based collaborative filtering approach. The term $\sum_{i_1 \in R_{v_k}}$ refers to the sum over all items i_1 that user v_k has already rated. In other words, the prediction takes into account the ratings that the user has given to other items i_1.

User-based CF methods, on the other hand, determine user similarity by analyzing the shared elements rated by both the target user and other users. They use these similarities and their corresponding weights to predict the user's rating for a particular item. The Pearson correlation coefficient is the most widely used method for computing similarity [178], as detailed in Equation 1.4.

$$\text{Sim}_{\text{user}}(v_k, v_l) = \frac{\sum_{i_1 \in I_{v_k} \cap I_{v_l}} (R_{v_k,i_1} - \bar{R}_{v_k})(R_{v_l,i_1} - \bar{R}_{v_l})}{\sqrt{\sum_{i_1 \in I_{v_k} \cap I_{v_l}} (R_{v_k,i_1} - \bar{R}_{v_k})^2} \cdot \sqrt{\sum_{i_1 \in I_{v_k} \cap I_{v_l}} (R_{v_l,i_1} - \bar{R}_{v_l})^2}} \tag{1.4}$$

The equation 1.5 will be used to predict user v_k's rating for item i_2.

$$\text{P}_{\text{user-based}}(v_k, i_2) = \bar{R}_{v_k} + \frac{\sum_{v_l \in N_{v_k}} (R_{v_l,i_2} - \bar{R}_{v_l}) \cdot \text{Sim}_{\text{user}}(v_k, v_l)}{\sum_{v_l \in N_{v_k}} |\text{Sim}_{\text{user}}(v_k, v_l)|} \tag{1.5}$$

The term $\bar{R}_{v_k}$ is the average rating of user v_k, while $\bar{R}_{v_1}$ refers to the average rating of user v_l. The equation considers the ratings provided by neighboring users v_l on item i_2, adjusting them based on their similarity to v_k, denoted as $\text{Sim}_{\text{user}}(v_k, v_l)$.

In summary, memory-based CF methods can be categorized based on which dimension of the user-item rating matrix they use to compute similarities. Item-based methods focus on identifying items rated similarly by multiple users, assuming that those items are likely to be similar if many users rate two items similarly. Conversely, user-based methods identify like-minded users, assuming that users with similar preferences will tend to like similar items. Memory-based CF methods are intuitive and straightforward to implement. However, they face significant challenges in scalability and often need help with sparse data, making it difficult to capture complex patterns or handle cold start scenarios effectively.

Model-based CF predicts unknown ratings by learning a model from the underlying data using machine learning or statistical techniques. These methods address some of the limitations of memory-based CF, such as data sparsity and scalability issues. Model-based methods can also be divided into two parts:

- **Traditional Model-based CF Techniques:** The traditional model-based CF Techniques are a set of algorithms used in RSs to predict user preferences based on past interactions. Unlike memory-based CF, traditional model-based CF techniques rely on mathematical models derived from domain knowledge or statistical assumptions [180]. They often use predefined models such as Bayesian classifiers [223], regression-based approaches [113], Matrix factorization method [198], and cluster-based CF [241], among others. Traditional models are often computationally efficient and require fewer resources, making them suitable for real-time recommendations on large datasets. Techniques like matrix factorization offer good interpretability, allowing for understanding latent features between users and items. However, traditional techniques struggle to capture complex non-linear relationships in data, especially when dealing with highly heterogeneous data. Moreover, these models often fail to provide practical recommendations for new users or items due to the lack of sufficient data for model training [201].

- **Deep Learning in Model-based CF Techniques:** Deep Learning (DL) has emerged as a powerful subset of model-based CF, leveraging complex neural network architectures to capture intricate patterns within user-item interaction data. Unlike traditional model-based methods, which often rely on simpler algorithms like matrix factorization or clustering, DL models can understand and model non-linear relationships within the data, making them particularly effective in large-scale and sparse environments [180].

In this context, Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and other advanced DL techniques have been integrated into CF to enhance recommendation accuracy. These models can automatically learn high-level feature representations from raw data, allowing them to capture subtle and complex interactions between users and items that traditional methods might miss [149]. For instance, DNNs can uncover latent factors from user behaviors, while CNNs can process sequential or spatial data, making them well-suited for tasks like personalized content recommendations [142].

By embedding DL within the model-based CF framework, these methods improve predictive performance and offer greater flexibility in handling diverse data types, such as text, images, and sequential user behavior. However, while DL models bring significant advantages, they also introduce challenges related to model interpretability, computational complexity, and the need for large datasets to train effectively [136].

The concept behind model-based CF methods is to develop a model from interaction data, extracting relevant patterns from the dataset, and then making recommendations without needing to access the entire dataset each time [114]. Model-based approaches offer better scalability for large and sparse datasets by identifying underlying patterns instead of relying on direct user or item comparisons. They can also address cold start issues by integrating supplementary data or adopting hybrid methods. However, while they can achieve high accuracy, model-based approaches often require more development time and specialized expertise due to the complexity of model design and tuning [247].

Memory-based CF algorithms calculate similarities between users or items using interaction data, a process known as memory-based filtering. In this approach, data is accessed in real time to compute these similarities, which are then directly used to generate recommendations. In contrast, model-based methods involve creating and adapting a predictive model to generate recommendations. Typically, the model is trained offline, and recommendations are generated online with near-instant results by applying the pre-trained model [171].

Content-Based Filtering Method

Content-based filtering (CBF) approaches recommend items similar to those the user has liked, or that match the user's attributes or preferences based on the item and user features [162]. In this approach, both items and sometimes users are represented by feature vectors, with features expressed as numeric or categorical values reflecting different aspects such as item characteristics (e.g., color, price) or user attributes (e.g., age, interests) [145]. Various similarity or dissimilarity measures can be applied to these feature vectors to evaluate how

similar two items are. Commonly used algorithms for this purpose include Euclidean [96] and cosine similarity, as shown in Equations 1.6 and 1.7.

$$\text{Dissim}_{\text{item}}(\mathbf{i_1}, \mathbf{i_2}) = \sqrt{\sum_{k=1}^{n} (i_{1k} - i_{2k})^2} = \|\mathbf{i_1} - \mathbf{i_2}\|_2 \tag{1.6}$$

$$\text{Sim}_{\text{item}}(\mathbf{i_1}, \mathbf{i_2}) = \frac{\sum_{k=1}^{n} i_{1k} \cdot i_{2k}}{\sqrt{\sum_{k=1}^{n} i_{1k}^2} \cdot \sqrt{\sum_{k=1}^{n} i_{2k}^2}} \tag{1.7}$$

Here, $\mathbf{i_1}$ and $\mathbf{i_2}$ are vectors containing n elements, where each element is indexed by k. $\text{Dissim}_{\text{item}}(\mathbf{i_1}, \mathbf{i_2})$ measures the distance between the vectors, and $\text{Sim}_{\text{item}}(\mathbf{i_1}, \mathbf{i_2})$ measures their similarity. The index k runs from 1 to n, ensuring that all elements of the vectors are considered in the calculations. The calculated (dis)similarity values are then used to generate a ranked list of recommended items, with items that are more similar to the user's preferences ranked higher. The quality of the recommendations depends heavily on the chosen features and how they are weighted. These methods, grounded in information retrieval principles, are versatile and can be applied across various domains to generate recommendations. However, CBF methods often struggle with cold start problems when there is insufficient content information available for new users or items.

In CBF methods, the system is likely to perform efficiently when items are effectively represented through a clear and structured set of features. However, the effectiveness of CBF methods depends on how well items and users are represented through relevant features. The system's performance can degrade if the feature set needs to be completed or relevant. Content-based RSs generate recommendations by analyzing the content of items, which can include textual information, images, or other structured data. They identify patterns within the content by extracting meaningful features that align with users' needs, preferences, and tastes [37].

While CBF and CF each have their strengths, they also face particular challenges in practical applications. To overcome these limitations, such as the cold start problem, data sparsity, and limited representation. RSs often employ Hybrid Filtering methods, combining CBF and CF features to deliver more accurate and personalized recommendations.

Hybrid Filtering Method

The Hybrid Filtering (HF) approach combines multiple filtering techniques to address the limitations of traditional RSs, such as the cold start problem, data sparsity, limited diversity, and scalability issues [64]. HF methods can incorporate elements like item content, user ratings, CBF, demographic data, and contextual information. Examples of hybrid systems include

weighted and switching models. In a weighted hybrid system, the final recommendation score is derived by integrating results from different recommendation techniques, while in a switching hybrid system, the most appropriate method is selected based on specific criteria [59]. For instance, Sharma et al. [174] introduced a hybrid book RS that merges CF and CBF, significantly improving recommendation accuracy, particularly in addressing cold start issues and data sparsity. Similarly, Alhijawi et al. [7] developed SemCF, a hybrid system combining semantic attributes and satisfaction-based user similarity metrics to generate more precise recommendations. Despite their advantages, hybrid methods may need help with new users or items with limited associated data, even when content-based features are incorporated [2].

An HF method can be used mathematically to formalize the combination of CF and CBF. The final recommendation score for user v_k on item i_m is calculated as follows:

$$\mathrm{P}_{\mathrm{hybrid}}(v_k, i_m) = \alpha \cdot \mathrm{P}_{\mathrm{CF}}(v_k, i_m) + (1-\alpha) \cdot \mathrm{P}_{\mathrm{CBF}}(v_k, i_m) \tag{1.8}$$

$\mathrm{P}_{\mathrm{hybrid}}(v_k, i_m)$ represents the overall recommendation score for user v_k for item i_m. $\mathrm{P}_{\mathrm{CF}}(v_k, i_m)$ and $\mathrm{P}_{\mathrm{CBF}}(v_k, i_m)$ represent the scores derived from CF and CBF respectively. $\alpha \in [0,1]$ can be adjusted based on the desired balance between CF and CBF methods. A higher α emphasizes CF, while a lower α gives more weight to CBF.

Although traditional RSs have been widely used and remain successful, they need help accurately capturing user preferences due to the cold start problem, data sparsity, and limited representation. These limitations can reduce the effectiveness of RSs and potentially lower user satisfaction [66]. As a result, improving RSs to better meet user needs is an urgent task. Researchers have increasingly turned to ML techniques, particularly DNNs, which excel at modeling complex, non-linear relationships in large-scale and sparse datasets, thereby enhancing the accuracy and relevance of recommendations.

1.1.2 Machine Learning-based Recommendation Systems

Machine Learning (ML) is a contemporary engineering approach that enables machines to learn patterns from data and make informed decisions, simulating certain cognitive functions to perform tasks effectively [199]. To improve recommendation performance and address the limitations of traditional RSs, such as the cold start problem and data sparsity, ML techniques, particularly DNNs, have been introduced and applied to RSs [251].

Numerous researchers have investigated the effectiveness of DNNs in RSs. For instance, a deep knowledge-aware network (DKN) using CNNs was developed to predict users' click probabilities in news recommendations. This model integrates semantic-level information from news articles and knowledge from an external knowledge graph, allowing it to capture

richer contextual and background information about user interests. By leveraging CNNs, DKN can effectively analyze the sequential patterns of words and entities within news content, thereby improving the prediction accuracy of user preferences [210]. Another example is a neural news recommendation system that leverages attention mechanisms to learn users' historical preferences. This system dynamically weighs different news articles based on how much attention they attract from users, allowing the model to prioritize more relevant articles. By applying attention mechanisms, the model can focus on the most critical parts of user interaction history, enabling it to capture short-term and long-term user interests for more accurate and personalized news recommendations [226].

Similarly, the BERT4Rec model uses the Bidirectional Encoder Representations from Transformers (BERT) architecture to analyze users' interests. BERT enables the system to model the contextual relationships between words in both directions, capturing deeper user intent and nuances in their interaction history. This allows the system to provide highly accurate recommendations by understanding individual items and the context in which they were consumed [191]. Table 1.1 provides a detailed comparison between traditional and ML-based RSs.

Table 1.1 compares traditional and ML-based RSs. Traditional RSs, such as CF, CBF, and HF, are generally more straightforward to implement and more interpretable. However, they face limitations, notably the cold start and data sparsity issues. ML-based RSs, leveraging DL and other advanced techniques, bring greater adaptability and can handle complex, large-scale data, dynamically adjusting to evolving user behaviors. Despite these benefits, ML-based systems introduce new challenges, particularly fairness and interpretability.

Both types of RSs are challenged by a fundamental demand for improved accuracy, leading to a key research question: *How can RS performance be enhanced by addressing the cold start and data sparsity issues, thereby providing more accurate recommendations and elevating user satisfaction?*

As accuracy improves, however, concerns arise over potential over-specialization, which can lead to filter bubbles. When users are primarily exposed to content that aligns with their preferences, their opportunities to explore diverse perspectives and new interests diminish, reinforcing existing biases [146]. While many ML-based RSs prioritize high precision and personalization [94, 173, 148], over-personalization can constrain user choice, creating a filter bubble effect that limits access to diverse viewpoints and content [14]. This brings forth another critical question: *How can we ensure users are exposed to diverse perspectives while maintaining high recommendation accuracy and preserving their freedom of choice?*

Promoting diverse information delivery is essential to breaking filter bubbles and fostering a fairer recommendation environment. By providing a broader spectrum of content and

Feature	**Traditional RSs**	**ML-based RSs**
Methods	CF CBF HF	DNNs CNNs BERT DL in Model-based CF
Advantages	Simple and easy to implement; Computationally efficient for moderately sized datasets; Interpretable results	Can capture complex, non-linear user-item interactions; Handles large-scale and sparse data effectively; Supports multi-modal data (e.g., text, images, user behavior); Can dynamically adapt to user behavior
Challenges	Cold Start Problem; Data Sparsity; Limited diversity in recommendations	High computational cost; Fairness issues; Potential for reinforcing filter bubbles
Recommendation Output	Based on user-item similarity; Primarily relies on historical data and item similarity;	More personalized and diverse recommendations; Supports dynamic learning from user behavior; Utilizes multi-modal data for enhanced and more diverse recommendations

Table 1.1: Comparison of Traditional and ML-based RSs

viewpoints, RSs can mitigate bias reinforcement and reduce the risk of isolating users in homogenous information spaces [39]. This highlights the need for responsible RSs, which aim not only to achieve high accuracy but also to address diversity and fairness, creating a more balanced and ethical recommendation approach.

This thesis aims to contribute to developing RSs that users can trust, ultimately enhancing user satisfaction and confidence in the system. In the following section, I will explore how Responsible RSs can tackle these challenges through critical principles, including accuracy, diversity, and fairness.

1.1.3 Responsible Recommendation Systems

To address the challenges of cold start, data sparsity, filter bubbles, and lack of fairness concerns raised by traditional and ML-based RSs, the concept of responsible RSs emerged. Responsible RSs focus on addressing the ethical implications of the decisions and actions made by intelligent autonomous systems. Responsible RSs focus on addressing the ethical implications of intelligent autonomous systems' decisions and actions, particularly ensuring fairness and diversity. In this context, "ethical" means preventing bias, providing equal access to a broad range of information, and ensuring users can make informed choices. This section will explore the critical components of responsible RSs, including precision, diversity, and fairness, which are critical to creating a balanced and ethical recommendation environment.

- **Diversity:** In responsible RSs, diversity is a crucial measure of their capacity to provide users with various options rather than solely focusing on accuracy. While accuracy is a core goal of any RS, responsible systems prioritize avoiding over-specialization by intentionally diversifying recommendations to prevent users from being confined to narrow content [34]. Controlling exposure based on predicted accuracy can limit users' choices, reducing their autonomy and reflecting a more restricted RS approach [259]. Therefore, a central mission of responsible RSs is to allow users to explore varied perspectives, increasing freedom of choice and countering the filter bubble effect.

- **Higher Precision:** DNNs, vital components of modern ML, have greatly enhanced recommendation accuracy across diverse fields, including autonomous driving, smart homes, and RSs [163]. While DNNs contribute positively to accuracy and user satisfaction by addressing limitations of traditional methods [176, 183, 206], responsible RSs must balance precision with diversity to ensure a responsible user experience. Although improving accuracy remains a primary aim, achieving a balance between precision and diversity is essential for a truly responsible RS that values user satisfaction and freedom of choice.

- **Fairness:** Fairness in RSs ensures that the algorithms do not discriminate against certain groups or individuals. For instance, RSs used in job recruiting platforms must ensure that no demographic groups are systematically disadvantaged due to biased historical data or algorithmic decisions. It is essential to design RSs that provide equitable access to information and opportunities for all users, regardless of their demographics or preferences. This includes ensuring a balanced representation of different perspectives and avoiding bias that could unfairly disadvantage any group of users [125].

Compared to traditional ML-based RSs, responsible RSs focus on being more user-centric platforms, prioritizing not only recommendation precision but also enhancing users' overall experience with the system [62]. Striking a balance between recommendation accuracy and diversity is critical in responsible RSs, as it directly influences user satisfaction and long-term engagement. While recommendation accuracy aims to predict items that users are likely to interact with, recommendation diversity ensures that users are exposed to a broader range of content, preventing monotony and keeping users engaged. Achieving a balance between these two aspects enhances user satisfaction and prevents issues like filter bubbles, where users are repeatedly exposed to a narrow range of content, limiting their exposure to diverse viewpoints and potentially reducing long-term engagement. By offering a wider variety of recommendations, responsible RSs can encourage users to explore new interests and maintain interaction with the platform, leading to a more sustainable system in the long run.

1.1.4 Summary

Responsible RSs focus on enhancing user satisfaction by prioritizing accuracy, diversity, and fairness. DNNs are adopted to mitigate limitations such as the cold start and data sparsity problems found in traditional recommendation methods, improve the precision of recommendations, and enhance user satisfaction. However, despite improvements in precision, research on how well these techniques mitigate long-standing challenges and balance precision with other factors, such as diversity, still needs to be improved.

Diversity ensures that recommendations encompass a broad range of content, enriching users' perspectives, protecting their freedom of choice, and preventing filter bubbles. This balanced information delivery is critical in exposing users to diverse viewpoints, preventing the reinforcement of pre-existing biases, and ensuring that users are not isolated in a homogeneous information environment. By providing a variety of content, RSs enhance user satisfaction and contribute to a fairer and more inclusive recommendation environment.

This thesis will provide a detailed analysis to compare traditional and responsible RSs, demonstrating the impact of responsible RSs on critical metrics such as accuracy, diversity, and user satisfaction. The findings will contribute significantly to future research on user-centered RSs. The third chapter of this thesis focuses on improving recommendation accuracy through advanced algorithms, while the fourth and fifth chapters explore the importance of diversity and fairness in responsible RSs. This comprehensive approach ensures that responsible RSs not only meet high standards of accuracy but also promote a diverse and fair user experience, ultimately fostering trust and satisfaction from users.

1.2 Research Questions

The development and implementation of RSs have significantly transformed user experiences across various online platforms. However, traditional and ML-based RSs face significant challenges, such as cold start problems, data sparsity, and filter bubble effects. These challenges can significantly reduce the effectiveness of recommendations and negatively impact user satisfaction. As the field continues to evolve, there is an increasing need to address these limitations by developing responsible RSs that improve accuracy and prioritize diversity and fairness. This thesis explores the potential of responsible RSs to improve recommendation performance and create balanced, fairer, more user-centric experiences. To guide this investigation, several research questions have been formulated. Each question is aligned with a specific model proposed in this thesis, as discussed in Section 1.4.

1. **Research Question 1:** How can the accuracy of RSs be improved while addressing the limitations of traditional RSs, such as the cold start problem and data sparsity, and maintaining responsible practices? This question is addressed by the proposed **Dual-Observation-based Recommendation (DOR)** model, which incorporates both local and global observational signals and leverages knowledge graphs to enhance predictive capability in sparse data environments. It also considers how to incorporate high-order and diverse information to support responsible recommendation practices.
    - **Sub-Research Question 1.1:** How can RSs incorporate diverse perspectives to mitigate over-specialization and filter bubbles?
    - **Sub-Research Question 1.2:** How can high-order relationships be leveraged in RSs to enhance both accuracy and diversity in recommendations?
2. **Research Question 2:** How can RSs enhance diversity and mitigate filter bubbles to improve user satisfaction and trust? This question is investigated through the

development of the **Responsible Graph-based Recommendation (RGRec)** model. RGRec identifies users affected by belief-based filter bubbles and applies belief nudging to gradually expose them to more diverse content, balancing engagement and content breadth.

- **Sub-Research Question 2.1:** How can we effectively identify and measure filter bubble effects in RSs?
- **Sub-Research Question 2.2:** What strategies can mitigate filter bubbles while maintaining user engagement?
- **Sub-Research Question 2.3:** Can a responsible recommendation approach balance content diversity and user preference effectively?

3. **Research Question 3:** How can RSs be designed to achieve responsible and balanced information delivery without altering existing algorithms? This question is explored through the **Agent-based Adaptive Information Neutralization (AAIN)** model. AAIN functions as an external layer that adjusts recommendation outputs based on sentiment balancing and adaptive neutrality principles, without changing the internal logic of existing RSs.

   - **Sub-Research Question 3.1:** How can we mitigate the effects of filter bubbles without modifying core recommendation algorithms?
   - **Sub-Research Question 3.2:** How can RSs balance contrasting sentiments to promote diverse information exposure?
   - **Sub-Research Question 3.3:** What model design can effectively implement adaptive neutrality and sentiment diversity to reduce biased perspectives?

1.3 Design of Study

This thesis aims to investigate the development and effectiveness of responsible RSs by addressing the identified research questions. The research follows a structured seven-step process, as depicted in Figure 1.3 and proposed by [121], ensuring a comprehensive and systematic investigation. This framework employs quantitative analysis to analyze large datasets, identify patterns, and draw significant conclusions, providing a detailed understanding of the research. This approach is particularly suitable for evaluating performance metrics such as accuracy, diversity, and fairness of RSs.

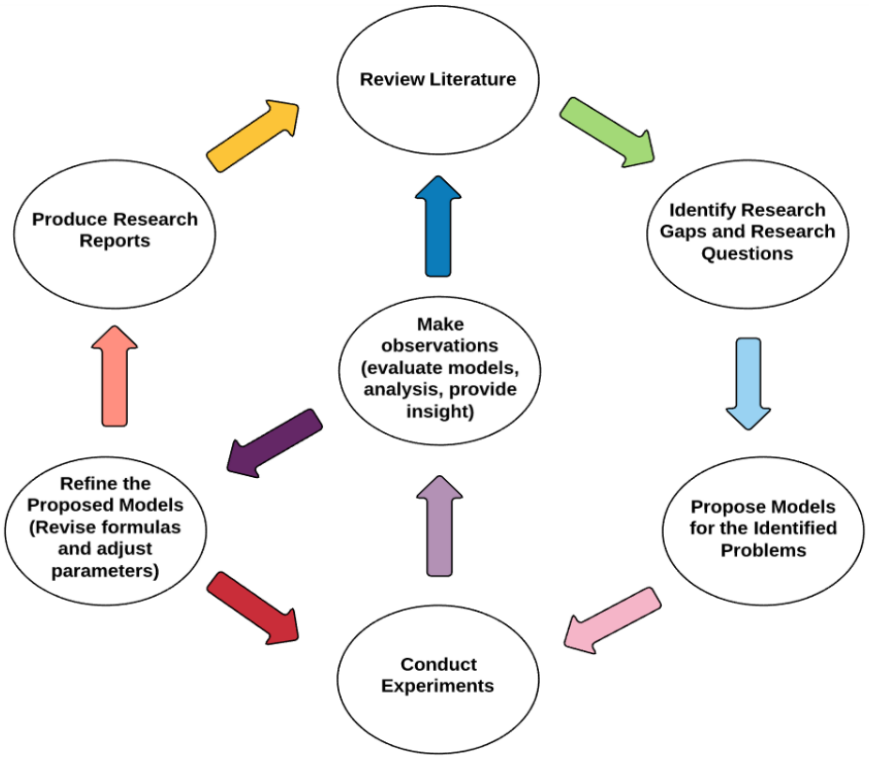


Fig. 1.3: Research Methodology Adopted in this Thesis

1.3.1 Research Methodology

In my Ph.D. study, the research follows a seven-step interactive methodology. The methodology is visualized in Figure 1.3, beginning with the literature review and concluding with the production of the research report. The methodology is outlined in detail below.

First, a thorough literature review is conducted to understand the current research in RSs, identify critical theories, methodologies, and findings, and highlight gaps requiring further research. Next, specific research questions are generated based on these gaps. These questions guide the study and ensure it addresses significant unresolved issues in RSs.

The third step is to propose models for the identified problems based on insights from the literature review, forming the foundation for empirical analysis. Next, experiments will be conducted, and quantitative data will be collected and analyzed. These experiments are built on the research questions and provide insights by analyzing experimental results. The study includes a refinement stage, where parameters are adjusted, formulas revised, and new variables incorporated to improve model effectiveness based on the analysis. Finally, the research reports and articles will be completed.

1.3.2 Evaluation Methods

Responsible RSs in this thesis will be evaluated based on several criteria, including recommendation accuracy, diversity, and fairness. Metrics such as precision, recall, diversity, fairness scores, and user satisfaction will be used to assess the system's performance across these criteria.

1. **System Perspective Metrics:** System Perspective Metrics evaluate the overall performance of RSs, focusing on technical and algorithmic aspects such as accuracy, diversity, and overall recommendation performance [112]. These metrics include:

   - **Recommendation Performance:**
     - **AUC (Area Under the ROC Curve):** AUC measures the probability that randomly chosen related items will rank higher than randomly chosen unrelated items. A higher AUC value indicates that the model can better distinguish between related and irrelevant items.
     - **MRR (Mean Reciprocal Rank):** MRR indicates the mean value of the reciprocal rankings of multiple query statements. It measures the effectiveness of a ranking system, with a higher MRR indicating a higher level of effectiveness.
     - **NDCG (Normalized Discounted Cumulative Gain):** NDCG measures the ranking quality of RSs. The principle of NDCG is that highly correlated products should rank higher than unrelated products. A higher NDCG value indicates a better ranking of related products. The NDCG@5 metric calculates the DCG of the first five recommendations while the NDCG@10 metric calculates the DCG of the first ten recommendations.
   - **Accuracy Metrics:**
     - **Hit Rate:** Measures whether at least one recommended item is accepted.
     - **Pre (Precision):** The ratio of accepted items in the recommendations.
   - **Diversity Metrics:**
     - **Coverage:** The sum of sentiments for a particular topic in the recommendations is divided by the sum of sentiments for that topic in the data pool and then averaged.
     - **RR (Repetition Rate):** Calculates the average sentiment redundancy for a specific topic. As the topic is τ_i, $RR = \frac{\sum_{i=1}^{n} RR_{\tau_i}}{n}$.
   - **Fairness Metrics:**

  - **Yin-Yang Neutralization Degree (best_diff):** Quantifies the fairness degree of RSs. It measures how well the system presents a neutral, balanced mix of content, ensuring that users are not overly exposed to one-sided perspectives. The detailed information on this metric is introduced in Chapter 4.

2. **User Perspective Metrics:** User Perspective Metrics assess the impact of RSs on users, focusing on user satisfaction and behavior. These metrics involve:

   - **Belief Network Diversity:** Evaluates the diversity of users' belief networks using the aspect coverage metric.
   - **Interest Evolution Analysis:** Analyzes how users' interests evolve in response to different recommendation strategies.

The above is a basic description of the various evaluation methods used in my study. The following sections provide more typical information.

1.4 Contributions of the Thesis

This thesis makes three major contributions by proposing novel models that directly address the challenges identified in the research questions.

1. **Enhancing Accuracy and Diversity Through Dual-Observation Mechanisms**

   - This thesis introduces the Dual-Observation-based Recommendation (DOR) model, which integrates both local and global observation networks to overcome traditional limitations like cold start and data sparsity. By leveraging high-order relationships and Knowledge Graphs (KGs), the model enhances accuracy while mitigating filter bubbles and over-specialization.

2. **Mitigating Filter Bubbles through Graph-based Detection and Nudging Strategies**

   - The Responsible Graph-based Recommendation framework (RGRec) is introduced, uniquely addressing filter bubbles by combining algorithmic and human-focused strategies. RGRec incorporates a Multi-faceted Reasoning-based Detection module (FBDetect) to identify filter-bubble-affected users and examines recommendation imbalances across diverse content. Additionally, RGRec leverages belief nudging techniques that guide users incrementally from highly preferred topics to less explored ones, broadening their perspectives and reducing ideological isolation. This iterative approach enables RGRec to present users with varied

content, fostering belief harmony and improving satisfaction by providing a more diverse recommendation environment without compromising user autonomy.

3. **Promoting Information Neutrality and Balanced Recommendations through Agent-Based Adaptive Neutralization**

    - This thesis presents the Agent-based Adaptive Information Neutrality (AAIN) model, which utilizes an innovative multi-agent framework inspired by the Yin-Yang theory to address the filter bubble phenomenon. By introducing the Adaptive Information Neutrality Agent (AINA) that applies Yin-Yang Neutralization Control (YYNC), the model counterbalances biases in preference-based recommendations. AAIN achieves neutrality by embedding diverse viewpoints within existing recommendation structures, providing users with a balanced exposure to information without altering core algorithms.

In summary, this thesis makes significant strides in advancing responsible RSs by developing three innovative models: the DOR, RGRec, and AAIN systems. These models incorporate advanced techniques such as DNNs, graph-based algorithms, and multi-agent systems to enhance recommendation accuracy while addressing common limitations in traditional and ML-based RSs. This research improves recommendation quality by emphasizing a balanced approach to accuracy, diversity, and fairness, exposing users to a more varied and equitable range of content. Altogether, these contributions provide practical solutions for building responsible, user-centered RSs.

1.5 Thesis Structures

The remainder of the thesis is constructed as follows:

1. **Chapter 2:** reviews several state-of-the-art studies and developments in RSs. It begins with the evolution of traditional RSs and then explores ML-based RSs, highlighting the use of ML and DL techniques to enhance recommendation accuracy and personalization. Key challenges such as cold start problems, data sparsity, filter bubble effects, and the balance between accuracy and diversity are discussed. Finally, the chapter introduces the concept of responsible RSs to ensure user-centric and trustworthy recommendations.

2. **Chapter 3:** introduces the Dual-Observation-based approach for Recommendation (DOR), a novel model designed to address the limitations of traditional RSs, such as

data sparsity, cold start, and the over-specialization nature of DL-based models. The DOR model leverages a dual observation mechanism, combining local and global observations to capture both user-specific and external perspectives, which enhances recommendation accuracy while promoting diversity. It addresses Research Question 1: How can the accuracy of RSs be improved while addressing the limitations of traditional RSs, such as the cold start problem and data sparsity, and maintaining responsible practices? This initial research was published in Applied Intelligence, titled "DOR: A novel dual-observation-based approach for recommendation systems" [214].

3. **Chapter 4:** focuses on achieving accuracy and diversity in RSs, emphasizing mitigating filter bubbles. It addresses Research Question 2: How can RSs enhance diversity and mitigate filter bubbles to improve user satisfaction and trust? The chapter explores strategies to incorporate diversity while maintaining accuracy and techniques to alleviate the negative impacts of existing filter bubbles on users. The final paper was published in IEEE Transactions on Artificial Intelligence, entitled "Nudging towards responsible recommendations: A graph-based approach to mitigate belief filter bubbles" [213].

4. **Chapter 5:** investigates the fairness of recommendation systems. It addresses Research Question 3: How can RSs be designed to achieve responsible and balanced information delivery without altering existing algorithms? The chapter explores methods to ensure balanced and fair information delivery and examines the impact of these factors on user satisfaction, trust, and acceptance. This work is currently under review by IEEE Transactions on Computational Social Systems.

5. **Chapter 6:** summarizes the thesis, emphasizing the motivation to address the limitations of traditional and ML-based RSs. It highlights the critical contributions of the developed models, the DOR system, the RGRec system, and the AAIN model, while acknowledging the limitations of the current research. Suggestions for future directions include further real-world testing, exploring new techniques to improve model performance, and understanding the long-term impacts on user behavior. Finally, the thesis underscores the importance of ethical, user-centric RSs that balance accuracy, diversity, and fairness, contributing to more effective and trustworthy environments.

Chapter 2

Literature Review

The literature review delves into the evolution and advancements of RSs, highlighting the significant changes from traditional to ML-based models. It explores the foundational principles of traditional RSs and their limitations, providing a detailed comparison between the two approaches. The review then introduces ML-based recommendation models, showcasing the role of ML techniques in enhancing personalization and accuracy while also discussing the disadvantages of these models. Additionally, it examines responsible RSs, which aim to tackle challenges related to diversity and user autonomy, ensuring that modern RSs are both practical and ethical. Finally, the review summarizes the existing limitations.

2.1 Traditional Recommendation Modelling

The origin of RSs can be traced back to the 1990s. With the popularization of the Internet, more users began to shop, read news, watch videos, and engage in other online activities. Helping users quickly find the content they are interested in from massive amounts of information became an important research topic [166].

2.1.1 CF Methods

One of the earliest RSs is CF, proposed by the Grouplens Research Group in 1992. The basic idea of CF is to predict items that users may be interested in based on their historical behaviors (such as ratings, browsing history, etc.) [87]. As mentioned before, CF can be divided into user- and item-based CF. Hernández and Gaudioso [53] proposed a framework that divides any RS into two distinct subsystems: one for guiding the user and another for delivering helpful or exciting items. The item-based CF method assumes that users who agreed in the past will also be interested in the future [158]. Sarwar et al. [169] first

introduced the basic item-based CF model, which calculates the similarity between items and predicts a user's rating for a target item based on the user's ratings of similar items. Unlike traditional CF, which focuses on finding similar users to generate recommendations, Linden et al. [127] extended the CF method further to build Amazon's item-to-item CF algorithm, which finds items similar to the ones a user has already purchased or rated and then recommends those similar items. The algorithm constructs a table of similar items by analyzing items that customers tend to purchase together. This is done through an iterative process of comparing item similarity based on shared customer purchases. Deshpande et al. [56] focused on refining and extending item-based CF methods and presented higher-order models that consider itemsets rather than just individual items, which can improve recommendation accuracy in specific contexts.

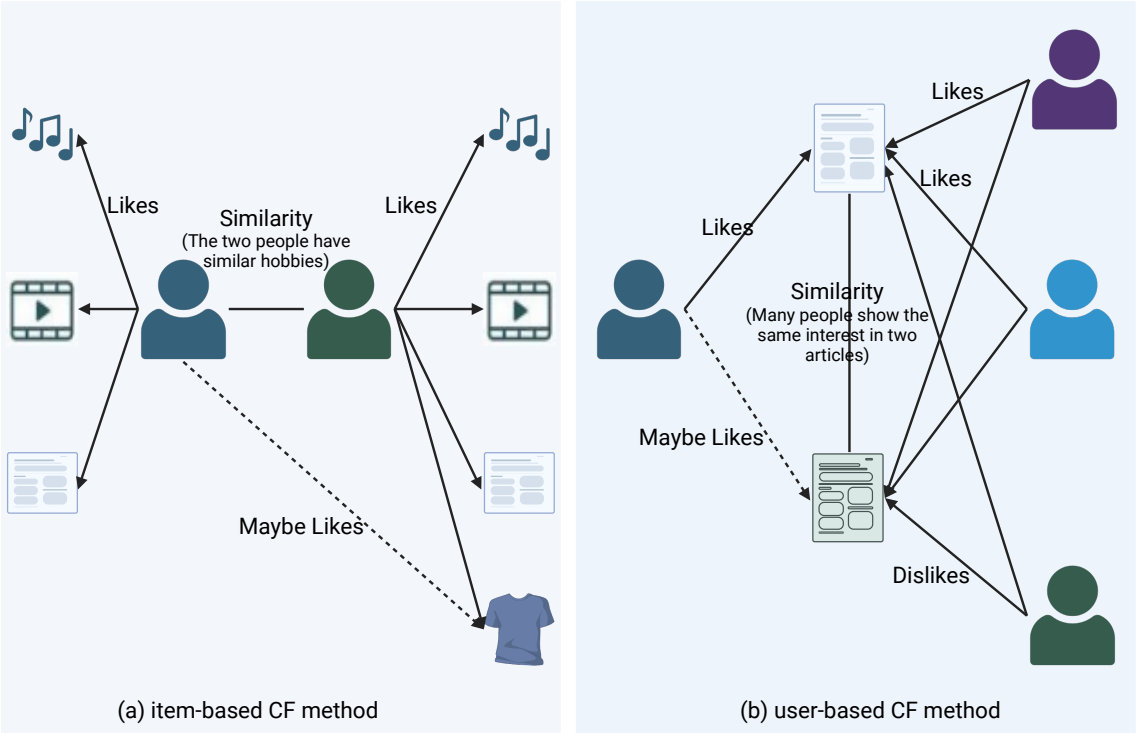


Fig. 2.1: Comparisons between item-based and user-based CF methods

Compared to the item-based CF method, the user-based CF method focuses on the relationships among users [144]. Nguyen et al. [144] proposed a cognitively inspired user-based CF framework that incorporates users' decision-making patterns and contextual influences into the similarity computation. Their model adjusts traditional similarity metrics by integrating cognitive load and time-aware decay functions, leading to more human-like

recommendation behavior. More detailed comparisons between item-based and user-based CF methods are displayed in Figure 2.1.

In addition to advancements in item-based CF, user-based CF methods have been extensively studied and improved over the years. However, most still rely on the availability of sufficient user-user interactions, making them vulnerable to data sparsity. The cognitive model by Nguyen et al., while more nuanced in capturing user behavior, still assumes access to rich contextual data and lacks mechanisms to address fairness. These limitations motivate later sections of this thesis, which explore how responsible models can better balance user preference modeling with diversity and equity considerations.

Sun et al. [192] introduced a novel user-based CF algorithm that incorporates data distribution, addressing the non-uniform distribution of user ratings. The algorithm significantly improves accuracy on sparse datasets while maintaining computational efficiency. This approach provides a more personalized recommendation considering the varying rating behaviors of users. However, it relies on the assumption of normal distribution in user ratings, which may limit its applicability across diverse datasets and require careful tuning to adapt to different data characteristics. Babu et al. [15] presented the implementation and evaluation of a user-based CF algorithm, highlighting its effectiveness in generating accurate recommendations by leveraging user similarity. Despite its simplicity and efficiency, the algorithm's performance depends on the number of neighbors selected and may face challenges such as high computational complexity and cold-start issues. Based on the previously mentioned, the CF method has been widely used in RSs, but it still faces various challenges, such as cold start, data sparsity, and scalability [174].

1. **Cold Start Issue:** It happens when a new user has not yet rated any items, preventing the system from recommending anything to them.

2. **Data Sparsity Issue:** This issue arises when there are more users than available ratings, as most users do not rate most items, resulting in a highly sparse user-item rating matrix. This situation typically occurs when the user-to-item ratio is high.

2.1.2 CBF Methods

Unlike analyzing user-item interactive records, the CBF method fixates on the item's features, such as a movie's types, description, or actors. The primary goal of CBF methods is to recommend items similar in content to those the user has previously liked. This is achieved by analyzing the attributes of items and comparing them to find similarities [8]. For example, if a user enjoys a particular movie, the system will recommend other movies with similar characteristics.

Mooney et al. [140] introduced LIBRA, a content-based book RS that uses a Naive Bayes text classifier to analyze the textual content of books rather than relying on user ratings, which are often sparse. By training the classifier on a user's previously rated books, the system predicts interest in new books based on textual similarity, making it especially effective for recommending less popular or newly added items. This work demonstrated that content-based machine learning approaches can yield accurate recommendations even in the absence of extensive user interaction data.

Similarly, Mak et al. [135] proposed INTIMATE, a web-based movie RS that employs text categorization techniques on movie synopses. Their evaluation compared different feature representations, such as bag-of-words and keyword-based, feature selection strategies, and classifiers. Results indicated that under conditions where the number of user ratings is low relative to the vocabulary size, the text-based approach can outperform traditional feature-driven methods.

Both systems leverage textual content to overcome cold-start scenarios, especially for new or unpopular items. However, they primarily focus on similarity and personalization, with limited consideration for recommendation diversity or fairness. These limitations highlight the need for RSs that not only address the cold start problem but also promote exposure to a broad spectrum of content, which serves as a key motivation for the models proposed in this thesis that seek to balance personalization with diversity.

CBF method has several advantages. It can provide highly relevant recommendations even with a small amount of user interaction data because it relies on item features rather than user behaviors. This makes it particularly useful for new users who still need to provide more interaction data to enable collaborative filtering. Additionally, the CBF method can explain its recommendations by highlighting the similar features between recommended items and items the user has liked in the past [97].

However, the CBF method also has limitations. One major challenge is the "over-specialization" problem, where the system may recommend items that are too similar to those the user has already seen, leading to a lack of diversity in recommendations. Furthermore, creating accurate and comprehensive feature representations for all items can be resource-intensive. This method also needs help with recommending items that do not have rich or detailed feature descriptions [186].

2.1.3 Hybrid Filtering Methods

HF methods are designed to combine the strengths of CF and CBF to provide more accurate and robust recommendations. CF relies on user-item interaction data to generate recommendations, while CBF focuses on items' intrinsic features. By integrating these two

approaches, HF methods aim to overcome the limitations inherent in each method when used independently.

CF is highly effective with ample user interaction data, as it can predict user preferences by finding patterns in historical user behavior. However, it needs help with the cold start problem and data sparsity. On the other hand, CBF can make recommendations based on item features even when interaction data is limited, making it particularly useful for new users or items. HF methods combine these approaches to handle situations where one method alone would be insufficient. For example, when a new item is introduced, CBF can generate initial recommendations based on its features, and as user interactions accumulate, CF becomes more effective.

Additionally, HF methods are more robust in dealing with data sparsity issues because they do not rely solely on user-item interaction data. By integrating user behavior and item content, hybrid methods can generate more personalized and accurate recommendations, even when user interaction data is sparse [120].

Ghazanfar et al. [79] proposed BoostedRDF, a cascading HF system that integrates item features, user ratings, and demographic data through a rule-based ensemble framework. This method enhances both precision and coverage, particularly in cold-start scenarios, by leveraging multiple information sources. However, the system's reliance on predefined rules and parameter tuning poses scalability challenges in large-scale environments. Similarly, Sharma et al. [174] introduced a hybrid RS that combines CF and CBF using an adaptive weighting mechanism. By dynamically balancing user interaction data and item content, the system improves accuracy in sparse data settings. Nevertheless, it depends heavily on the availability of high-quality side information and incurs higher computational costs.

Both approaches demonstrate the effectiveness of hybrid strategies in addressing cold start and sparsity. However, their performance is constrained by the complexity of the system and the requirements of the data. These limitations motivate the need for hybrid recommendation models in this thesis that are not only accurate but also scalable and adaptable.

Despite their success, HF methods come with challenges. Implementing them can be more complex than using a single approach, as they require careful tuning and balancing of multiple models. Furthermore, combining CF and CBF models often necessitates significant computational resources for training and real-time recommendation generation. This can result in higher operational costs and complexity. However, HF methods reduce the limitations of traditional CF and CBF models by addressing the cold start problem, managing data sparsity more effectively, and better handling complex user preferences [195].

Metrics such as precision, recall, and F1 score are commonly used to evaluate the performance of HF methods. These metrics help assess recommendations' accuracy, relevance, and

overall effectiveness [12, 81]. By leveraging the best CF and CBF approaches, hybrid methods provide a more flexible and scalable solution to recommendation challenges, although they require careful optimization and increased computational resources.

2.1.4 Summary of Traditional Recommendation Modelling

Traditional recommendation modeling, encompassing methods like CF and CBF, laid the foundation for modern RSs. These techniques addressed the initial challenge of helping users find relevant content in vast amounts of information online. Through user- and item-based approaches, CF methods leverage historical user behavior to predict future interests. Meanwhile, CBF methods focus on the intrinsic features of items to find similar content for users. HF methods combine the strengths of both CF and CBF, offering a more robust solution to handle data sparsity and cold start problems. However, traditional methods still need to improve in cold start, data sparsity, and managing complex user preferences. This necessitates the integration of more advanced technologies. Consequently, the evolution of RSs has led to the need for ML techniques, which promise to enhance recommendation accuracy and personalization further. The following section explores these ML-based recommendation models and their impact on the field.

2.2 Machine Learning-based Recommendation Modelling

ML has revolutionized the landscape of RSs, providing advanced techniques to overcome the limitations of traditional methods, such as cold start problems and data sparsity [161]. Traditional RSs, including CF and CBF, rely heavily on user-item interaction data or item features, but they often need help when such data is limited. In contrast, ML-based RSs offer more personalized and accurate recommendations by leveraging sophisticated models that can handle sparse data, uncover latent patterns, and better understand complex user preferences [248].

As a subset of AI, ML introduces techniques such as k-nearest neighbors (k-NN), support vector machines (SVM), and matrix factorization. These methods differ fundamentally from traditional CF and CBF approaches regarding computational complexity, data requirements, and flexibility in capturing hidden relationships within the data [114].

For instance, matrix factorization, a widely used ML technique, decomposes the user-item interaction matrix into latent factors, enabling the prediction of missing entries by capturing underlying preference patterns between users and items [207]. It typically employs algorithms such as stochastic gradient descent (SGD) or alternating least squares (ALS)

to iteratively optimize the latent embeddings. Unlike traditional CF methods that rely on explicit similarities, matrix factorization reveals hidden dimensions that can improve recommendation performance, especially under sparse data conditions. However, the method is computationally intensive and requires careful regularization to avoid overfitting, especially in large-scale systems.

Similarly, k-NN recommends items by identifying the top-k most similar users or items based on distance metrics like cosine similarity or Euclidean distance. Its simplicity and transparency make it appealing for small datasets or real-time recommendation scenarios. However, its pairwise similarity computations become a bottleneck when scaling to larger datasets, and its direct reliance on user-item interaction data limits its effectiveness under data sparsity. ML-based variants address this by incorporating dimensionality reduction or learning adaptive similarity metrics [19].

SVM, a supervised learning algorithm, is commonly used in RSs for classification and ranking tasks. Its ability to handle high-dimensional and non-linear data makes it suitable for modeling complex user preferences [156]. Joachims [104] introduced SVMrank, which focuses on learning personalized item ranking rather than pointwise scores, offering a more tailored recommendation. SVMs are especially effective when contextual features such as user demographics or temporal behaviors are available. However, they face scalability issues due to their high training complexity and the need for extensive hyperparameter tuning [109].

While each of these ML techniques offers distinct advantages, such as matrix factorization for latent preference modeling, k-NN for simplicity, and SVM for interpretability and non-linearity, they also come with notable drawbacks in scalability, data dependency, and engineering effort. Moreover, most of these models primarily focus on accuracy and personalization, with limited consideration for exposure diversity, fairness, or the long-term societal impacts of recommendations.

These limitations suggest the need for more responsible recommendation strategies, which not only ensure predictive performance but also address issues such as over-specialization, filter bubbles, and unequal information exposure. These concerns motivate the models proposed in this thesis, which aim to balance predictive accuracy with broader ethical and user-centric considerations.

2.3 Knowledge Graph-based Recommendation Modelling

KGs are structured and semantic representations of knowledge that are becoming increasingly significant in RSs. A KG consists of entities (also known as nodes) and relationships (also known as edges) between those entities. Each entity represents a distinct concept, object, or

item, such as a book, movie, or person, while the relationships describe how these entities are interconnected [33]. For instance, in a movie RS, entities might include movies, directors, actors, and genres, with edges representing relationships like "directed by", "acted in", or "belongs to genre".

Unlike traditional data structures used in RSs, such as user-item matrices (CF and CBF methods), KGs incorporate rich contextual information and semantic relationships between items. Traditional structures often rely on simple, flat representations where the connections between items or users are non-existent or minimally defined, focusing primarily on direct interactions like ratings or clicks. In contrast, KGs incorporate a deeper layer of context by explicitly modeling the relationships and attributes that define these interactions, thus providing a more nuanced understanding of user preferences and item similarities [117, 33].

This ability to represent complex interconnections is a crucial difference between KGs and traditional data structures. Traditional RSs often need help dealing with sparse data or understanding the more profound, non-obvious relationships between items. KGs address these limitations by enabling RS to leverage semantic information, leading to more accurate and personalized recommendations. By understanding the underlying relationships between items, KGs help to uncover connections that might not be immediately apparent, thus offering recommendations that are not only relevant but also enriched with contextual meaning [43, 60].

2.3.1 Knowledge Graphs and Recommendation Systems

KGs have become increasingly important in the RSs field due to their ability to address several critical challenges, such as cold start, data sparsity, and over-specification. Traditional RSs often need more data about a user or an item, leading to less accurate or even irrelevant recommendations. KGs mitigate these issues by leveraging their rich, interconnected data, enabling the system to infer relationships and preferences even when direct interaction data is sparse [85].

Moreover, KGs allow RSs to move beyond simple item-to-item correlations, which are often the basis of traditional recommendation methods. Understanding the complex relationships between entities, such as how two seemingly unrelated movies might share a common theme, director, or genre KGs, enables more context-aware recommendations. This ability to capture and utilize semantic relationships makes KGs particularly valuable for enhancing the accuracy and relevance of recommendations and providing explanations that users can understand and trust [85, 181].

2.3.2 Knowledge Graphs with Machine Learning

Integrating KGs with ML models has opened new avenues for enhancing RSs. KGs can generate embedding representations of entities and relationships that capture the rich contextual information in the graph. These embeddings can then be fed into ML algorithms to improve the accuracy and relevance of recommendations [71].

One approach to integrating KGs with ML is through graph neural networks (GNNs), designed to operate directly on the graph structure. GNNs can learn from the relationships and features of nodes in the graph, effectively capturing the dependencies between entities. For instance, Graph Convolutional Networks (GCNs), a type of GNN, have been successfully applied in recommendation tasks to propagate information across the graph, allowing the model to consider both direct and indirect relationships in its recommendations [212].

Several successful implementations of KGs in RSs have demonstrated their potential to improve performance significantly. A notable example is the Deep Knowledge-Aware Network (DKN), which integrates KGs into the recommendation process using a combination of Convolutional Neural Networks (CNNs) and KGs to provide more contextually aware recommendations. The DKN uses embeddings derived from the KG to capture the semantic relationships between entities, which are then processed by the CNN to make final recommendations [210]. Another example is the KG Attention Network (KGAT), which leverages the attention mechanism within a GNN to selectively focus on the most relevant parts of the KG when making recommendations. This approach allows the model to prioritize certain relationships or entities, leading to more personalized and accurate recommendations [220].

In addition, the Description Enhanced KG (DEKR) model represents a significant advancement by jointly incorporating structural relationships and textual descriptions into the KG [32]. DEKR employs a GNN to propagate entity-level signals across the KG while integrating a text-aware collaborative filtering component that captures the semantic relevance of item descriptions. By combining structural and textual modalities through joint learning, DEKR improves representation quality and enhances recommendation accuracy. Experimental results show that DEKR outperforms prior KG-based methods, such as KGAT and DKN, particularly in tasks such as click-through-rate prediction and top-K recommendations.

Compared to earlier models that rely solely on structural embeddings, DEKR demonstrates the effectiveness of incorporating unstructured information to enrich entity semantics. This integration is especially beneficial for items with limited relational connections, offering an implicit solution to the cold start problem. The model's dual-channel design underscores the potential of multi-source data fusion in overcoming sparsity and improving both personalization and content coverage.

These insights support the thesis's direction toward designing models that go beyond structural reasoning and incorporate additional context, such as textual information or user-centric signals, to enhance diversity in recommendation outcomes.

2.3.3 Advantages of Using Knowledge Graphs in RSs

KGs significantly enhance recommendation accuracy by leveraging the semantic relationships between entities, which traditional RSs might overlook. For example, in a movie RS, a KG can help recommend films based on user preferences for genres or actors and more nuanced relationships, such as a director's style or the themes explored in the films. This allows the system to make connections that might not be immediately apparent in a traditional user-item matrix, leading to more precise and personalized recommendations.

The cold start problem, where new users or items have little interaction data, is a common challenge in traditional RSs. KGs help alleviate this issue by using the entity relationships in the graph to recommend items, even when direct interaction data is unavailable [115]. For instance, if a new movie is added to the system, the KG can recommend it to users based on its connections to other movies (such as shared genre, director, or actors) that the user has shown interest in.

KGs also promote diversity in recommendations by linking items across different but related categories. This is particularly useful in avoiding the over-specialization problem, where users are only shown items similar to their past preferences. By understanding the broader relationships between entities, KGs can introduce users to a wider variety of content, encouraging exploration and discovery and ultimately enhancing the overall user experience [143].

2.3.4 Summary of KG-based Recommendation Modelling

KGs can potentially transform RSs by improving accuracy and diversity. By capturing and leveraging the rich semantic relationships between entities, KGs enable RSs to go beyond simple correlations and provide more contextually aware and personalized recommendations. Additionally, KGs address key challenges like the cold start problem and data sparsity, making them a powerful tool for enhancing user satisfaction and engagement. Looking ahead, the future of KG-based RSs lies in the continued integration of KGs with advanced ML techniques to further enhance the capabilities of KGs in providing accurate and diverse recommendations.

2.4 Responsible Recommendation Modelling

The rapid development of ML in RSs has significantly enhanced personalization and accuracy. However, these advancements have also introduced new challenges regarding diversity and user autonomy, while there is a continual need to improve recommendation precision and user satisfaction further. Responsible recommendations address these challenges by ensuring that RSs are practical but also ethical and trustworthy [55]. Responsible ML is the core technology of RSs, which involves developing and deploying ML-based systems ethically and accountable. This approach considers the broader implications of ML technologies, ensuring they align with societal values and protect user rights. In the context of RSs, responsible ML emphasizes the need for higher accuracy, diversity, and user autonomy [58].

2.4.1 Social Impacts of Recommendation Systems and Diversity

RSs have become vital to the digital landscape, influencing how users discover information, products, and content. While RSs have greatly enhanced the personalization of content, they also raise concerns about their broader social impacts. Among the most urgent concerns are filter bubbles and echo chambers, which can limit users' exposure to diverse perspectives and information [75].

Filter Bubble and Echo Chamber

The concept of a filter bubble was popularized by Eli Pariser in 2011 [189], referring to the personalized environments that are created when RSs continuously show users content that aligns with their existing preferences and behaviors. This personalization can lead to a narrowing of perspectives, causing users to be increasingly exposed to content that reinforces their existing beliefs while being shielded from contrasting viewpoints.

While related to filter bubbles, Echo chambers refer to a distinct yet complementary phenomenon. An echo chamber emerges at the group or community level, where information, ideas, or beliefs are amplified and reinforced through repetition among members of a closed group. In such an environment, opposing viewpoints are often excluded or marginalized, creating a space where only similar perspectives are shared and discussed [47]. Unlike filter bubbles, which are primarily the result of algorithmic filtering by RSs that personalize content for individuals, echo chambers can develop both algorithmically and socially. Users may naturally gravitate toward communities or groups that align with their pre-existing beliefs, further reinforcing their views. In these environments, dissenting opinions are

often suppressed or absent altogether, leading to increased polarization and a deepening entrenchment of beliefs [61].

In a filter bubble, users are repeatedly exposed to similar content, which exaggerates their current opinions and potentially deepens their views by limiting exposure to diverse perspectives. Similarly, echo chambers magnify this effect by creating self-reinforcing feedback loops, where users are surrounded by content that consistently echoes their existing beliefs. While filter bubbles are typically algorithmically generated and focus on personalized content for individuals, echo chambers can form organically within social groups and communities, often reinforced by social dynamics and selective content exposure. RSs, particularly those driven by engagement metrics, can worsen these issues by prioritizing content that is more likely to result in user interaction, such as likes, shares, or clicks, over content that might challenge the user's views or broaden their perspectives [132].

Addressing the social impacts of RSs, particularly filter bubbles and echo chambers, is a crucial component of responsible RS development. Responsible RSs aim to break the filter bubble and dismantle echo chambers by balancing personalization with diversity, ensuring users are exposed to a broader range of content, including differing viewpoints. This approach benefits individual users by fostering a more balanced and well-rounded view, and it also serves the larger goal of reducing societal polarization.

Several research efforts have focused on mitigating the negative impacts of RSs on content diversity and user exposure. For example, Belavadi et al. [21] and Gharahighehi et al. [77] proposed diversity-aware algorithms that introduce a broader range of content into users' recommendation feeds. These approaches incorporate diversity metrics such as intra-list diversity, coverage, or novelty scores into the ranking process to optimize not only for accuracy but also for exposure variety. Such models aim to reduce the formation of echo chambers and mitigate the intensity of filter bubbles.

Beyond algorithmic solutions, responsible RSs can incorporate user feedback mechanisms to give individuals control over the diversity of their content. Dalecke and Karlsen [50], for instance, developed an interactive RS interface where users can specify preferences for content from different genres, viewpoints, or topics. This explicit feedback loop allows users to dynamically guide the recommendation process, potentially improving transparency and trust while combating over-specialization.

While algorithm-driven methods offer automated control over diversity, user-driven approaches empower individuals to shape their content exposure. The choice between the two depends on context, but both reflect a broader trend toward responsible personalization.

These studies emphasize the importance of balancing personalization with exposure to diverse content, aligning with this thesis's objective of designing models that integrate algorithmic fairness, user agency, and diversity optimization into the recommendation process.

Accuracy and Diversity

In responsible RSs, accuracy and diversity are essential for fostering user trust, satisfaction, and long-term engagement. Accurate recommendations ensure that users receive content closely aligned with their preferences and needs, thereby enhancing the perceived value of the system. This accuracy allows RSs to provide highly personalized experiences, increasing user trust and satisfaction. However, an overemphasis on accuracy can lead to over-specialization, commonly known as the "filter bubble" effect, where users are only exposed to content that reinforces their existing beliefs and preferences. This effect limits their exposure to diverse perspectives and experiences, creating an isolated, narrow view of the world [218].

To mitigate these issues, incorporating diversity into RSs is equally crucial. Diversity ensures that users are exposed to a broad range of content, promoting exploration and discovering new interests. By presenting a variety of viewpoints and content types, diversity prevents biases and encourages users to explore beyond their usual preferences, fostering more dynamic and engaging interactions with the system. A diverse set of recommendations can significantly enhance the user experience and contribute to a more balanced and inclusive digital ecosystem.

Balancing accuracy with diversity is vital to avoiding the pitfalls of over-specific recommendations while maintaining relevance. This balance is crucial for sustaining user satisfaction and long-term engagement. By combining accuracy with diversity, RSs can encourage users to explore new content while providing relevant recommendations that meet their needs [57]. Achieving this balance requires integrating diversity metrics into recommendation algorithms. These metrics ensure that recommended content spans various topics and categories, thus enhancing user exposure to diverse content without sacrificing relevance.

Future directions in responsible RS development include the integration of ML frameworks that consider the long-term social impacts of recommendation algorithms. These frameworks could guide the design of RSs that are accurate and engaging and foster a more inclusive and diverse digital system. For example, diversity metrics such as the Generalist-Specialist Score (GS-Score) [11], Individual Diversity [35], and User-oriented Group Fairness (UGF) [125] can be implemented to measure and enhance diversity in various dimensions. The choice of specific diversity metrics depends on the research objectives, but they are vital

in understanding and improving the overall user experience by promoting a broader range of content exposure [54].

In conclusion, while RSs have revolutionized how users interact with content online, they also carry significant social responsibilities. By addressing the challenges of filter bubbles and echo chambers, responsible RSs have the potential to enhance recommendation quality, promote diversity, and ultimately create a more informed and connected society. A well-balanced RS, incorporating both accuracy and diversity, can meet individual user needs and contribute to a healthier, more pluralistic digital environment.

2.4.2 Fairness in Recommendation Systems

Fairness in RSs is a critical concern, as algorithmic personalization often risks reinforcing users' existing beliefs and biases [197]. Traditional recommendation algorithms, particularly those centered around user preferences, can inadvertently lead to "filter bubbles", limiting exposure to diverse viewpoints. Such narrowing of perspectives can create ideological echo chambers, contributing to social and political polarization [13, 88]. Therefore, fairness in recommendations goes beyond accuracy and involves a commitment to delivering diverse, balanced content that mitigates these biases.

To address these issues, many RSs incorporate fairness by balancing content diversity while aligning with user preferences. Diversity-promoting techniques, such as re-ranking and clustering, are frequently adopted to introduce varying perspectives into recommendation lists [154, 25]. Re-ranking methods often adjust an initial recommendation list by optimizing for fairness-aware objectives (such as increasing exposure of underrepresented items), while clustering-based strategies aim to segment users or items to ensure group-wise fairness across different subpopulations.

Tang et al. [197] proposed the Unbiased Fairness-Aware Recommendation (UFAR) framework, which directly addresses bias in observational data using a range-based Inverse Propensity Score (IPS) and soft ranking metrics. This combination enables UFAR to achieve more equitable and accurate Top-N recommendations across multiple datasets without heavily compromising relevance.

Complementing these algorithmic approaches, Sonboli et al. [187] emphasized the user-side perspective by exploring how fairness and transparency are perceived in RSs. Through semi-structured interviews, they uncovered that many users lack awareness of how recommendations are generated and what fairness entails. They argue that transparency should be treated not just as an informative layer, but also as an educational tool that builds user trust and fosters more equitable human-AI interaction.

While algorithmic and user-centric fairness methods serve different roles, they are often complementary. Algorithmic fairness modifies system behavior, whereas transparency interventions support user comprehension and trust. However, achieving fairness remains challenging due to the deep-rooted biases in preference-based models. Many RSs prioritize relevance, which often results in reinforcing user-specific patterns and narrowing exposure [152, 44]. Adjusting algorithmic output for fairness frequently requires re-engineering recommendation pipelines, which may not be practical for systems that prioritize efficiency or legacy stability.

In this thesis, fairness is redefined as the delivery of balanced and unbiased information by neutralizing overly specific content exposures, rather than by altering the core recommendation algorithm. This approach, grounded in the principles of neutrality and exposure diversity, seeks to ensure that users engage with content in a way that is equitable, transparent, and minimally intrusive to existing system architectures. By embedding fairness through post-hoc neutralization and exposure shaping, the proposed models provide a scalable and interpretable alternative to structural re-design.

2.5 Summary

This literature review outlines the progression of RSs, highlighting key advancements and emerging considerations. Initially, RSs helped users filter massive amounts of online content using techniques such as CF and CBF. CF predicts user preferences by analyzing historical user-item interactions, while CBF recommends items by focusing on item features and identifying similarities with previously liked items. However, CF, CBF, and HF have limitations in handling new users or items, known as the “cold start” problem, and sparse datasets.

The advent of ML introduced more sophisticated models to tackle these challenges. ML techniques like matrix factorization, clustering, and DL models enable RSs to uncover hidden patterns within data, offering more personalized and accurate recommendations. Matrix factorization, for instance, improves RS accuracy by identifying latent factors in user-item interactions. Clustering techniques and DL models, such as neural networks, enable RSs to handle large datasets and complex relationships between users and items. However, these ML models are computationally intensive and often operate as “black boxes”, where the decision-making process is complicated for users to interpret. Additionally, the emphasis on personalization can lead to over-specialization and filter bubble effects, where recommendations become too narrow, continuously reinforcing a user’s preferences and limiting exposure to diverse content.

KGs offer another significant enhancement to RSs by representing items and their relationships in structured graphs, capturing contextual and semantic relationships between items. By linking related entities, KGs allow RSs to provide more context-aware and relevant recommendations, leveraging connections beyond user-item interactions. When combined with ML, KGs help RSs manage sparse data more effectively, addressing the cold start problem and allowing for more nuanced recommendations.

With the increased complexity of RSs comes a focus on responsible recommendation approaches to align RS outputs with ethical considerations such as diversity and fairness. Responsible RSs aim to balance the accuracy of recommendations with the need for fairness, minimizing risks like filter bubbles, which can isolate users within narrow perspectives. By incorporating user feedback mechanisms, responsible RSs promote more balanced content exposure and build trust by clarifying recommendation logic. As a result, a table 2.1 is provided for a comparative overview of mainstream recommendation models, highlighting their core advantages, limitations, and ideal application domains. This summary supports subsequent model selection and evaluation in the context of RS design.

Table 2.1: Comparison of Recommendation System Approaches

Model Type	**Representative Models**	**Key Advantages**	**Main Limitations**	**Typical Application Scenarios**
Memory-based CF	UserKNN, ItemKNN	Simple, interpretable, and responsive to real-time updates	Struggles with data sparsity and scalability	Small-scale systems with dense user-item interactions
Model-based CF	SVD, ALS, SVD++	High accuracy, captures latent preferences	Cold start issues, limited interpretability	E-commerce, video streaming platforms
CBF	TF-IDF, LDA, Deep Content Models	Handles new item recommendations well; interpretable	Limited in modeling user intent shifts	Text, music, academic recommendations
Hybrid RSs	Feature-weighted Hybrid, Switching Hybrid, Ensemble	Combines strengths of multiple models, robust performance	Increased complexity, hyperparameter tuning required	Multi-source recommendation, news platforms
Traditional ML Models	SVM, Random Forest, XGBoost	Easy to train and interpret; effective on structured data	Limited representational power; relies on manual feature engineering	Binary classification tasks, rating prediction
DL Models	NCF, Wide & Deep, DeepFM	Strong representation capacity; models complex nonlinear patterns	High training cost, lacks transparency	Large-scale systems, social media
KG-based Recommenders	RippleNet, KGCN, KGAT	Incorporates semantics and context; enables cross-domain reasoning	Costly to construct and maintain KGs; computationally intensive	Academic resource recommendation, semantic search
Responsible Recommendation Models	UFAR, GS-Score, UGF	Promote fairness and diversity; mitigate societal risks	Balancing accuracy and responsibility is complex	News feeds, social platforms, educational content

Chapter 3

Enhancing Responsible Recommendations with a Dual-Observation Mechanism

3.1 Introduction

The rapid expansion of web and mobile applications grants users easy access to vast amounts of global information. However, this information abundance presents a challenge: how to provide accurate recommendations that match user interests while avoiding a narrow, single perspective. Responsible RSs aim to address this by fostering exposure to diverse perspectives while also aligning content with user preferences for enhanced personalization and accuracy [227].

Traditional recommendation methods, such as CBF, CF, and HF, have been widely used to tailor content based on user preferences [215]. Despite their success, these approaches often face data sparsity and cold start issues [209]. To address these limitations, researchers have incorporated additional data sources and DNNs to improve user interest modeling. Yet, existing methods may struggle to capture nuanced propagation patterns effectively [116].

Capturing high-order propagation information is essential for enriching user behavior features and providing contextual insights into user interests. While DNNs can enhance accuracy in modeling these interests, an overemphasis on user-specific preferences can lead to over-specialized RSs, often referred to as "filter bubbles," which limit user exposure to diverse content [132].

To overcome these challenges, it is critical to incorporate diverse information into the recommendation process. By utilizing KGs, which leverage global information, RSs can

expose users to a broader range of content and knowledge. This approach not only improves recommendation accuracy but also mitigates filter bubble risks by ensuring that users encounter a variety of perspectives [116].

Recently, integrating KGs with DNNs has become a popular method to capture both low- and high-order relationships in RSs [210, 212]. For example, Wang et al. combine DNNs for modeling low-order relationships with GNNs to capture high-order interactions [212]. This combined approach improves recommendation diversity and addresses filter bubbles while maintaining accuracy.

To address these needs, I propose a novel approach called the Dual-Observation-based approach for Recommendation (DOR). The DOR framework integrates local and global observation networks to represent user interests comprehensively. Specifically, the local observation network is a content-based model that captures user-specific belief networks through textual analysis, while the global observation network leverages KGs to explore interactions between user beliefs and diverse external information. This dual observation mechanism broadens users' content exposure, reducing filter bubble risks while delivering accurate, personalized recommendations.

- Different readers may focus on different aspects of the same article for different reasons. For example, an article headline such as *"New Zealand fully reopens to the world in August: Ardern"* may attract readers who are interested in the economic impacts of the border reopening, as well as those who are interested in the real estate market. Therefore, it is important to consider the various belief networks of readers in the recommendation process.

- One-sided or incomplete observations may not be sufficient to form a satisfactory user belief system. A user's belief system is shaped by all of the information that this user has encountered. This information should include all relevant contextual semantics and be revised as the user's beliefs change. Therefore, it is essential to leverage dual observation mechanisms to deeply explore the semantic information in the textual information and analyze the influences of each article on the user's belief network.

- Highly accurate RSs that focus narrowly on users' past interests risk over-specification, leading to filter bubbles where users are repeatedly exposed to similar perspectives. Such systems may fail to introduce new and diverse content that could expand users' knowledge and perspectives. To counter this effect, incorporating diversity into recommendation algorithms is crucial. By broadening users' content exposure, RSs can mitigate the risks of filter bubbles, ensuring that users are presented with a variety of information that encourages a more responsible recommendation behavior.

To address the three challenging issues mentioned above, in this chapter, I propose a novel approach called the Dual-Observation-based approach for Recommendation (DOR). The DOR architecture incorporates a local and global observation-based interest extraction and construction model, which simultaneously distills and learns local and global users' interest representations. By integrating dual observation neural networks, the DOR approach can endow each piece of user behavior with deeper meaning. Specifically, the local observation network is a content-based learning model that imports textual inputs, forming the users' belief networks. On the other hand, the global observation network leverages KGs to explore the mutual influence between users' belief networks and various extrinsic information sources. By incorporating KG fusion, the DOR approach enhances users' exposure to more diverse content, reducing the risk of filter bubbles. By considering both local and global observations enriched by KGs, the DOR approach provides more accurate, personalized, and diverse recommendations, ensuring a more responsible user experience.

The contributions of this chapter are summarized as follows.

- I propose a novel Dual-Observation-based approach for Recommendation (DOR), a model that leverages both local and global observation networks. This dualistic approach enables a deeper semantic exploration of target items, refining the construction of user preferences. The DOR approach bridges the gap between item content and users' belief networks while incorporating external knowledge to avoid over-specialization and mitigate the risk of filter bubbles.

- I advance the state-of-the-art by integrating low and high-order relation expressions within our model. This combination offers a powerful solution to common issues like data sparsity and cold start. By balancing low and high-order expressions, the DOR approach ensures robust and reliable recommendations, even in situations with sparse user interaction data or new users, while enhancing diversity in content exposure.

- For high-order relation expression, two advanced methods are proposed. First, I introduce hybrid-domain information from external sources using KG fusion, which improves the generalization ability of the model and reduces the limitations of data sparsity. Second, I propose an attention-based graph-enhanced global contextual (AGGC) model to extend attention by considering the global context of the KG and enriching user preferences with diverse and relevant information.

- I validate the superior performance of the DOR model through extensive experiments on real-world datasets. The experimental results demonstrate that DOR significantly outperforms existing baselines across multiple evaluation metrics. This empirical validation highlights the practical utility and effectiveness of our approach.

The remainder of this chapter is organized as follows: Section 3.2 reviews literature related to this study. Section 3.3 introduces some preliminary concepts of our research and the problem description. Subsequently, the details of our proposed dual-observation-based recommendation system are given in Section 3.4. Section 3.5 describes the experimental settings and demonstrates the results of the experiments. Finally, I conclude this research work and point out future directions in Section 3.6.

3.2 Related Works

3.2.1 Feature-based Recommendation

RSs have long relied on feature-based techniques, with CF being a key approach. CF is widely adopted for capturing user-item interactions [216]. However, CF algorithms often suffer from cold-start and sparsity issues, which limit performance when user data is scarce [220]. CBF was introduced to address these issues by calculating similarities based on content features [97]. Wynne et al., for example, applied CBF to model a fake news detection system [237]. However, CBF models often require extensive manual feature engineering, which can be time-consuming and resource-intensive [229].

To enhance these models, researchers have integrated additional factual data, leading to hybrid methods that combine CF and CBF techniques. For instance, Sharma et al. proposed a hybrid recommendation model that successfully applies CF and CBF to book recommendations [174]. Wu et al. leveraged social influence to enhance users' acceptance of recommended incentives [234]. While hybrid approaches have improved accuracy and alleviated cold-start issues, they still fall short in capturing complex user-item interactions and adapting to changing user preferences. To address these challenges, DL techniques have emerged, offering more robust solutions by enabling non-linear transformations and sequence modeling [4].

3.2.2 Deep Learning-based Recommendation

DL techniques have been widely adopted and applied to various applications, including recommender systems [3, 167, 256]. Zhang et al. summarize several classical DNNs for recommendations, including Multilayer Perceptron (MLP), Autoencoder (AE), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Restricted Boltzmann Machine (RBM), Neural Autoregressive Distribution Estimation (NADE), Adversarial Network (AN), Attentional Models (AM), and Deep Reinforcement Learning (DRL) [253]. These approaches offer several benefits for RSs, such as the ability to model complex user-item interactions

through nonlinear transformation, learn rich item representations through representation learning, model sequential user behavior through sequence modeling (SM), and increase flexibility.

Numerous studies leverage DL techniques for recommendation tasks[253]. Hui et al. use an embedding-based model to represent items [94], while Wang et al. propose a time-aware method, which is built on the RNN method to mimic user behaviors [202]. Tang et al. propose a dynamic graph-based RS that can capture users' evolving preferences toward items over time [196]. Wu et al. devise a deep reinforcement learning-based method to recommend incentives for promoting users' beneficial behaviors [233]. Zhu et al. develop an RS based on the RNN model called the Deep Attention Network (DAN) model, showing the importance of the RNN method in fully exploring users' historical sequential features [260]. Chen et al. design a co-occurrence CNN that considers both user-item and item-item interactions [41]. Guo et al. utilize DL techniques to address the limitations of previous social recommendation research, such as insufficient robust data management and overly specific preferences [86]. These DL-based approaches can automatically discover information-rich item expressions without the need for extensive manual processes and often provide a better understanding of item content than feature-based techniques.

However, DL-based systems are prone to over-specification, meaning they often provide highly tailored recommendations based on a user's existing preferences. While this can enhance relevance, it may lead to a narrowing of content exposure, as users are continuously presented with similar types of content. This over-specialization can result in a lack of diversity in recommendations and increase the risk of filter bubbles, where users are exposed to only a limited range of perspectives or topics. Addressing these limitations, improving model interpretability, and ensuring diversity continue to be key priorities for DL-based RSs [26].

3.2.3 KG-based and Responsible Recommendation Systems

Several researchers incorporate KG techniques and DL skills into diverse tasks. Shi et al. adopt a concern graph and graph-based presentation skills to improve the public concern detection effectiveness and achieve a satisfactory result [179]. In recent years, researchers have been focusing on integrating KGs into RSs, resulting in several successful approaches. For example, Sun et al. demonstrate the effectiveness of KGs in improving recommendation satisfaction by using both KGs and DNNs to obtain item representations [193]. Zhang et al. design a graph-based context-aware RS with a KG to analyze and predict users' behavior [249]. Wang et al. propose a KG-based recommendation model that learns movie representations at a high-order level, achieving satisfactory results [211]. Ma et al. incorporate hybrid

information, such as news categories and six different types of behaviors, to construct the user's behavior graph and further demand news diversity [134]. Fan et al. also integrate KGs into the recommendation model, using Graph Neural Networks (GNNs) to learn features from duplicate user-user and user-item graphs [70].

In text-based scenarios, KGs have been adopted to extract semantic meaning. For example, Sheu et al. apply a KG to the news domain, focusing on exploring the contextual features of news to represent users' reading interests over a short period, where Graph Convolutional Networks (GCNs) are used to embed contextual information [177]. Wang et al. incorporate a KG into a news RS for news content engineering, using TransE to represent news entities and pre-trained Word2Vec to express word embeddings. The findings indicate that the utilization of a KG has a profound influence on the effectiveness of recommendations [210].

In recent years, responsible RSs have gained increasing attention as researchers strive to improve both the relevance and ethical dimensions of recommendations, such as reducing filter bubbles and over-specialization. KGs have become a valuable tool in achieving these goals by introducing broader contextual knowledge and diversifying recommendation outputs. KGs are particularly beneficial in addressing challenges like content diversity and the limitation of single-domain user interest, all of which are critical in responsible recommendation contexts.

Several studies highlight the potential of KGs in diversifying recommendation content and reducing user exposure to narrow viewpoints, effectively mitigating "filter bubble" risks. For example, Daniels et al. demonstrate how integrating KGs with DNNs can enhance user satisfaction by enriching item representations with external contextual information, which supports exposure to a broader range of topics and perspectives [51]. Similarly, Lops et al. proposed an end-to-end framework (ClayRS) designed to enhance replicability in knowledge-aware RSs. This framework provides a complete pipeline for building, evaluating, and experimenting with advanced knowledge representations, aiming to standardize and advance responsible recommendation practices [131].

These studies indicate that KGs serve not only as a means to improve recommendation accuracy but also as a tool to uphold responsible recommendation principles by promoting diversity. By integrating KG data with recommendation algorithms, RSs can provide a varied user experience, making KGs a vital component in the development of responsible, user-centered RSs.

3.2.4 Attention-based RSs using Deep Neural Networks

In recent years, researchers have integrated attention mechanisms into RSs to enhance performance. Attention serves as a technique that enables models to identify the crucial

elements of input data that are pivotal for decision-making [100]. The attention method distinguishes the importance of data by learning patterns within it and using those patterns to prioritize certain parts of the data when making decisions, enabling the model to focus more heavily on the most relevant features and improving its decision-making capability [216]. On top of that, attention mechanisms enable personalized recommendations by allowing the model to focus on relevant information and automatically extract relevant information, improving the understanding of the item's content [239].

Jung et al. use an In-and-Out Attention flow framework in a dialogue RS [105], while Zhu et al. develop an attention-based DNN news RS [260]. Wu et al. consider diverse news information in their proposed recommendation model and include an attention mechanism [228]. They represent users' interests from word-level expressions and incorporate category embeddings into the news embeddings. Similarly, Li et al. adopt a similar method, using an attention-based deep neural network RS in various scenarios [118]. Duan et al. leverage the CNN model and multi-attention mechanism for KG-based recommendation task [63], highlighting the non-trivial influence of relations in contextual representation learning. These studies demonstrate the positive impact of attention networks on recommendation research.

However, few studies consider the interaction between the user's knowledge system and input information from a macro perspective and the context-rich semantic expression of input information.

3.2.5 Summary

Existing RSs often struggle to capture the nuanced, bidirectional influences between users' belief networks and a wide range of global information sources, leading to a limited representation of user preferences and interactions. Moreover, challenges like the cold-start problem and the tendency toward over-specialization reduce the system's ability to recommend diverse content, contributing to potential filter bubbles. The proposed Dual-Observation Recommendation (DOR) model addresses these limitations by introducing a dual-observation approach that considers both local and global dimensions. Through local observation, DOR captures detailed contextual semantic information from users' direct interactions. Meanwhile, global observation incorporates diverse external information sources, enhancing the system's understanding of user interests in broader contexts.

The DOR model leverages both low-order and high-order relation mechanisms to refine interest representations, achieving a balanced view of user preferences. This dual mechanism not only alleviates cold-start limitations by enriching the recommendation space but also enhances content diversity, effectively mitigating filter bubble risks. By aligning users' belief networks with both immediate and extended information sources, DOR promotes

transparency and fosters a responsible recommendation environment that better serves user interests and encourages exposure to diverse content.

3.3 Preliminary

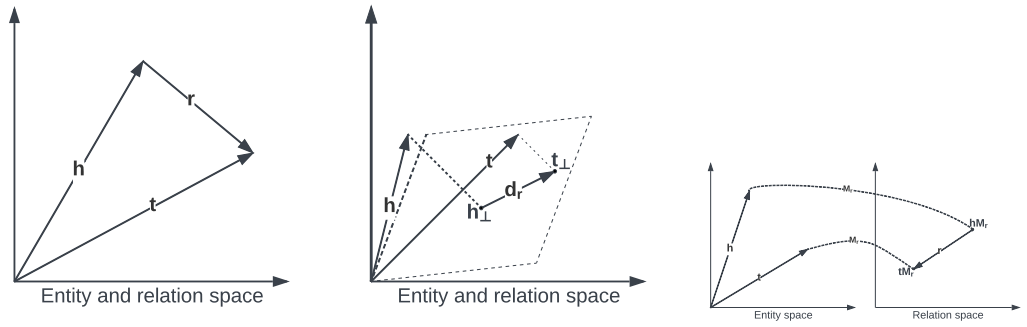
This section aims to provide an introduction to several fundamental concepts that are essential for this research. These concepts include KG embedding and dual observations. Subsequently, I will formulate the problem within the context of the current setting.

3.3.1 Knowledge Graph Embedding

KGs have been widely studied from many perspectives, including representation and modeling, knowledge identification, knowledge fusion, and knowledge retrieval and reasoning [42]. Benefiting from the power of KGs, incorporating KGs into RSs has become popular in recent years, as it can significantly improve the performance of recommendations [243]. In general, a KG consists of several Resource Description Framework (RDF) triples, where each RDF triple contains a head entity h, a relation r, and a tail entity t [150]. To effectively derive and utilize information on entities and relations in the KG, it is necessary to represent them as low-dimensional vectors in a continuous space, i.e., embeddings. These embeddings can be used for subsequent tasks, such as link prediction, entity classification, and knowledge base completion [252].

There are several approaches for learning KG embeddings, such as neural network-based models [212], semantic matching models [258], and translation-based models [257]. The translation-based models have proved their effectiveness and efficiency in representing entities and relations, and they have the complexity of space and time that scales linearly with the dimensionality of entities and relations embedding space [129]. Furthermore, neural network-based embedding models and semantic matching models usually suffer from the limitation of data sparsity and over-simplify [139]. Hence, I select three widely used translation-based models (i.e., TransE, TransH, and TransR) to represent KG triples into low-dimensional embeddings in the proposed DOR system. The details of these three models are as follows:

- **TransE** [107] is a translation-based model that learns low-dimensional embeddings of entities and represents relationships as translations in the embedding space. The objective of TransE is to minimize the distance between vectors $\mathbf{h}+\mathbf{l}$ and $\mathbf{t}$ if the triple (h,r,t) holds, or to maximize the distance conversely, as described in Figure 3.1a. Accordingly, the scoring function of TransE can be represented using Equation



(a) TransE (b) TransH (c) TransR

Fig. 3.1: Simplified illustrations of entities and relations in TransE, TransH, and TransR

3.1. Although TransE can effectively handle 1-to-1 relations in the KG, it has flaws in managing 1-to-N, N-to-1, and N-to-N relations.

$$f_r(\mathbf{h},\mathbf{t}) = -\parallel \mathbf{h}+\mathbf{r}-\mathbf{t} \parallel_2^2 \tag{3.1}$$

- **TransH** [102] overcomes the problems of TransE in modeling 1-to-N, N-to-1, and N-to-N relations by enabling entities to have distributed representations in different relations. Specifically, as described in Figure 3.1b, TransH first introduces a hyperplane $\mathbf{w}_r$ to represent a specific relation r and positions the translation vector $\mathbf{d}_r$ in the hyperplane. Then it projects $\mathbf{h}$ and $\mathbf{t}$ to the hyperplane $\mathbf{w}_r$, denoted by $\mathbf{h}_\perp$ and $\mathbf{t}_\perp$, respectively. The expectation is that $\mathbf{h}_\perp + \mathbf{d}_r$ approaches $\mathbf{t}_\perp$ if (h,r,t) holds. Equation 3.2 formulates the scoring function of TransH.

$$f_r(\mathbf{h},\mathbf{t}) = -\parallel \mathbf{h}_\perp+\mathbf{d_r}-\mathbf{t}_\perp \parallel_2^2 \tag{3.2}$$

  where $\mathbf{w}_r$ is the normal vector, $\mathbf{h}_\perp = \mathbf{h} - \mathbf{w}_r^\top \mathbf{h}\mathbf{w}_r$, and $\mathbf{t}_\perp = \mathbf{t} - \mathbf{w}_r^\top \mathbf{t}\mathbf{w}_r$. Through this mechanism, TransH enables entities to have diverse roles in different relations.

- **TransR** [69] further improves the embedding performance by modeling entities and relations into distinct embedding spaces because entities and relations are completely different objects. To perform the translation in such settings, TransR sets entities embeddings as $\mathbf{h},\mathbf{t} \in \mathbb{R}^k$ and the relation embedding as $\mathbf{r} \in \mathbb{R}^d$ for any triple (h,r,t), where $k \neq d$. Meanwhile, a projection matrix $\mathbf{M}_r \in \mathbb{R}^{k\times d}$ is used to project entities from the entity embedding space into the relation embedding space, as shown in Figure

3.1c. The scoring function of TransR is described in Equation 3.3.

$$f_r(\mathbf{h},\mathbf{t}) = -\parallel \mathbf{h}\mathbf{M}_r + \mathbf{r} - \mathbf{t}\mathbf{M}_r \parallel_2^2 \tag{3.3}$$

3.3.2 Dual Observations

In the current setting, the dual observation aims to access and refine the user's reading preference by considering both the focus of the text-based information and the user's belief network. It consists of local observation and global observation.

The local mechanism combines low-order relation representations with high-order relation expressions, which helps to alleviate data sparsity and the cold-start problem and allows for the exploration of hybrid-domain and more meaningful information for the user. The global observation mechanism focuses on the continual influence of each piece of information on the user. Every time a user reads a piece of information, their belief network is connected to the primary semantic information and historical knowledge of the information due to the mutual influence between each word and the influence between the user's belief network and the information. Different users may show different levels of attention to the same information. Hence, observing local textual features and the interaction between the user's belief network and the information is essential for extracting the user's current knowledge system.

Dual observation differs from dual attention, which generally refers to using two separate attention mechanisms in a single model, where the two attention mechanisms can be used independently or in combination to attend to different aspects of the input data [147]. For example, a model with dual attention normally uses one attention mechanism to focus on user information and another attention mechanism to focus on item information when making recommendations [100]. By contrast, dual observation emphasizes the vital information of the article and the user's belief network.

Figure 3.2 demonstrates how readers perceive, understand, and integrate information. Users form their belief system essentially through two observation mechanisms. The first is the observation of the local attention of the textual information, which entirely extracts the textual feature (the word with red color, with darker red representing more critical). When users read this article, they will analyze and perceive the information based on their existing belief networks, describing their prior knowledge and experiences. The pure blue user belief network represents the user's prior knowledge. The second observation mechanism is global observation, which observes the mutual influence between the user's prior knowledge and the outside source. The final affected result (mixed color user belief networks) is transferred back to the user. It can be seen that when a person reads a text, they receive the article's

information and incorporate it into their existing belief networks. It is important to adopt dual observation mechanisms to retrieve users' preferences accurately.

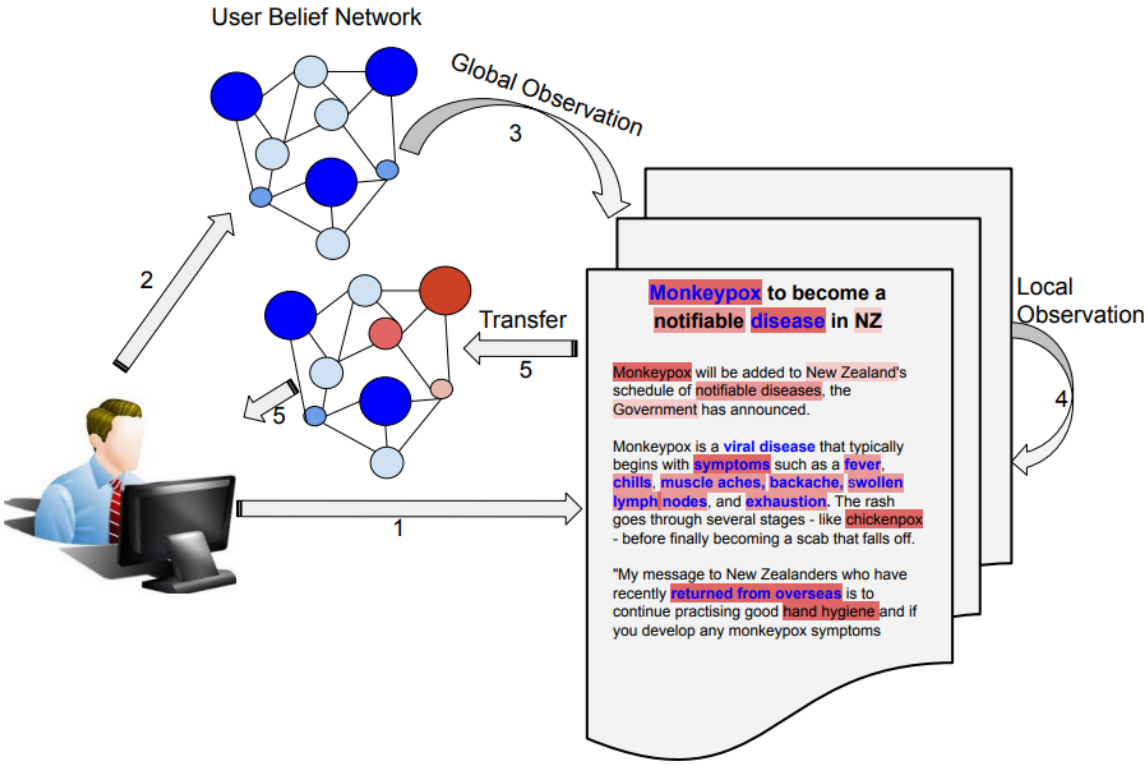


Fig. 3.2: A typical process of how readers perceive and understand information

3.3.3 Problem Definition

This research introduces a novel approach to responsible personalized recommendations by integrating users' belief networks with diverse contextual and semantic features. Through a dual-observation mechanism, the model captures both user-specific and external perspectives, promoting personalization while mitigating risks like over-specialization and filter bubbles, thereby ensuring a more responsible recommendation experience. I formally define the reading behavior of a reader r as $b_{r,t} \in B_r$, where B_r represents the list of historical reading behaviors of reader r. Each reading behavior $b_{r,t}$ is represented as a four-tuple, i.e., $b_{r,t} = (reader_{id}, item_{id}, t, l)$. Here, $reader_{id}$ denotes the reader's unique identifier, $item_{id}$ refers to the identifier of the textual item associated with the behavior, t represents the timestamp, and l indicates the label that denotes whether the item was clicked or not. Each item in the dataset comprises a title or overview, which can be represented as a sequence of words $[w_1, w_2, \ldots, w_m]$. m indicates the length of the news. Our objective, with a focus on

responsible recommendation, is to predict the likelihood of reader r selecting a candidate item while ensuring diverse content exposure and calculating the corresponding click probability.

3.4 Dual-Observation based Recommendation

In this section, I comprehensively explain the Dual-Observation based Recommendation (DOR) model, starting with an overview of its architecture. Subsequently, I delve into the details of the dual-observation modules, particularly emphasizing the local and global observation mechanisms. This thorough exploration will shed light on the key components and workings of the DOR model.

3.4.1 DOR Architecture

The overall architecture of the proposed DOR model is demonstrated in Figure 3.3. The model primarily comprises two main modules: the local and global observation mechanisms.

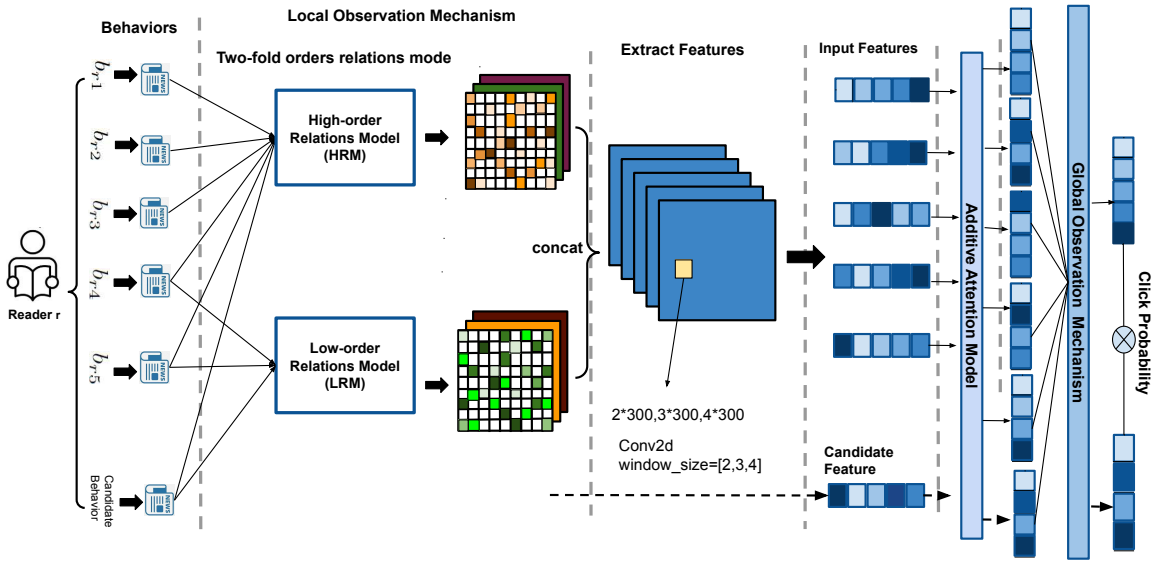


Fig. 3.3: The overall architecture of DOR

As illustrated in Figure 3.3, a reader r's reading histories $B_r = \{b_{r1}, b_{r2}, b_{r3}, b_{r4}, b_{r5}\}$ are fed into the DOR system. The local observation mechanism processes these inputs by utilizing the high-order relation model (HRM) and the low-order relation model (LRM). These two relations models learn textual features from contextual and content perspectives. The HRM module incorporates global knowledge from KGs to overcome the limitations of information sparsity and learn contextual representations at a high-order level. Meanwhile,

the LRM module is used within the local observation mechanism to extract rich features to obtain lexical-level representations. The feature extractor concatenates and then trains the high-low bipartite representations of target textual information. Using an additive attention model, the local observation mechanism observes the local weights to differentiate between different attentions among entities and words for the user. On the other hand, the global observation mechanism delves into the deep relationships between the user's evolving beliefs and the input information. As the user receives new information, the corresponding belief is continuously changing. The global observation mechanism examines the significant mutual influence between the user's belief and the input information. Lastly, similarity calculation calculates the probability of the user clicking on the candidate input.

The details of the local and global observation mechanisms are given in the following subsections.

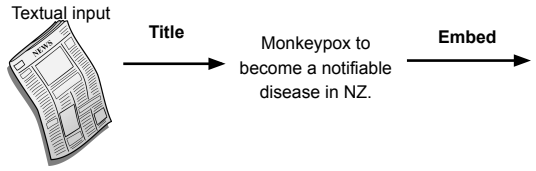
3.4.2 High-Order and Low-Order Relations

The local observation module concentrates on the content and contextual information of each item. It aims to identify the inherent characteristics of articles by utilizing both high and low-level representations. The readers' perception of the information is partially influenced by the main idea of the reading materials, where different words are assigned varying levels of importance.

As depicted in Figure 3.3, the local observation mechanism includes two essential models, i.e., the low-order relations model (LRM) and the model of the high-order relation (HRM). The LRM captures the lexical-level representations of textual inputs, and the HRM incorporates global knowledge from KGs to extract contextual representations at a high-order level. These two models work in tandem to provide a comprehensive understanding of the characteristics of articles. Next, the LRM and HRM are introduced with examples.

The low-order relation module (LRM) captures the intra-article word relationships in the current configuration. Each word in the article is encoded as a vector, enabling the depiction of connections between words. The LRM module aims to derive lexical-level representations of articles by considering these word relationships, which are then used for further analysis. This mechanism allows the DOR system to obtain local information from articles, leveraging it for recommendations. Figure 3.4 illustrates an example of the LRM module, where the input is titled "Monkeypox to become a notifiable disease in NZ". The title undergoes encoding using a pre-trained word embedding model, resulting in its representations.

On the other hand, high-order relations describe the connections between entities mentioned in the article that are not explicitly mentioned in the text but can be inferred through



	1	2	3	4	5	...	300
Monkeypox	0.27742	-0.7913	0.31968	-0.4115	-0.086		-0.3865
to	0.2919	0.1706	-0.2661	0.3915	-0.162		0.668
become	0.27015	-0.4515	0.4352	0.6811	-0.2031		0.2599
a	0.3645	-0.249	0.2327	0.2923	-0.0276		-0.653
notifiable	-0.0639	-0.7455	-0.1217	-0.1342	0.03801		0.1611
disease	-0.3944	-0.3002	-0.1547	-0.09	-0.1456		-0.288
in	0.6521	-0.3036	0.40041	0.17576	0.2646		0.02964
NZ	-0.027	0.2924	-0.2705	0.14258	0.3578		0.5467

Fig. 3.4: An Example of Low-order Relations Model (LRM)

additional context or global knowledge. To identify these connections, it uses techniques such as knowledge distillation and entity linking [244], which align entities present in the textual content with pre-defined entities in a global KG. In this research, I use Wikidata [1] as the global KG, which is a vast repository of structured data from the real world and can involve rich content to users. To provide a clearer understanding of the High-order Relation Module (HRM), an example is depicted in Figure 3.5. The HRM module utilizes triple distillation techniques to extract subjects, objects, and predictions from textual inputs. In the example, the subject "Abandoned Theme Parks" and the object "Nostalgia" are identified as entities within the input. Additionally, the prediction "Explore" is also extracted, representing the connection between the two entities. However, due to potential limitations in extraction, the keywords of the textual input are also considered as entities in our model and aligned with corresponding entities in a sub-KG. Consequently, closely related entities (represented as green nodes in the graph) associated with the input keywords (represented as yellow nodes in the graph) are further extracted from the sub-graph and integrated into existing triples (represented as blue nodes in the graph). By incorporating global knowledge, the HRM module enhances contextual understanding of the article, which in turn aids the recommendation process.

3.4.3 Graph Feature Learning

I employ graph feature learning to extract entity neighbors and relations from a graph, which is then utilized in a graph-based feature learning mechanism known as the Attention-based Graph-enhanced Global Contextual (AGGC) model. The AGGC model leverages

[1] https://www.wikidata.org/

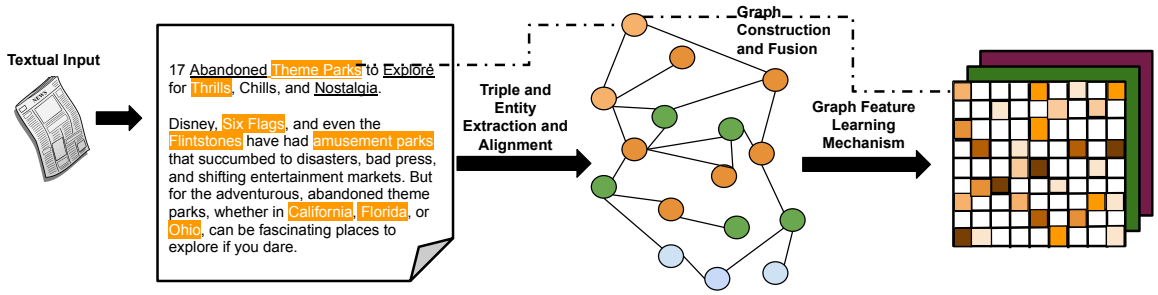


Fig. 3.5: An Example of the High-order Relations Model (HRM)

contextual information, such as multiple-head neighbor expressions and their associated relation representations to discover embeddings for each node. It further improves input representations by incorporating a global KG and considering contextual information within the graph. The GFLM highlights the varying importance of each node, enabling a better understanding of the significance of individual entities.

The proposed AGGC model combines deep neural networks with advanced attention mechanisms to capture intricate relationships within the data. It builds upon the success of existing graph attention models, such as the Graph Attention Network (GAT), but introduces key enhancements that significantly improve performance and versatility. Different from traditional graph convolutional networks (GCN) that aggregate information from immediate neighbors, and GAT that selectively attends to neighbors using attention mechanisms, the AGGC extends the concept of attention by incorporating contextual information into the attention mechanism. It considers the global information or context of the graph when assigning attention weights to neighbors. By integrating local and global information, the AGGC allows nodes to attend to neighbors that are relevant in a broader context, capturing more holistic information from the graph and enhancing the representation learning process.

The AGGC model architecture consists of multiple layers of graph convolutional operations (GCNConv), followed by non-linear activations. The attention layer is introduced to compute attention weights based on the output of each GCNConv layer. These attention weights are then applied to the GCNConv outputs using element-wise multiplication and aggregation. Finally, a linear layer is used for further transformation before producing the final context embeddings. To train the AGGC model, an optimizer (e.g., Adam) is employed to minimize the defined loss function between the predicted context embeddings and the target values. The model parameters are updated iteratively over a specified number of epochs.

The AGGC model significantly advances capturing complex relationships within graph data by incorporating global contextual information through attention mechanisms. Its unique

ability to attend to relevant neighbors in a broader context (hybrid domains) enhances the representation learning process and enables a more comprehensive understanding of each entity's significance. Thus, the AGGC model stands as a valuable and innovative approach for graph-based learning tasks. The detailed algorithm of the AGGC is described as follows.

The AGGC presents a novel neural network architecture that deals with graph-structured information, which reuses the concept of "local" and "global" and fixates on contextual information from a global perceptive. The input of the user belief graph with a set of nodes represented by E and a set of edges expressed by R. Each node e_i in this user network is associated with a feature vector $\mathbf{e}_i \in \mathbb{R}^d$, where d is the dimension of the vector feature. The purpose of the AGGC model is to calculate a new representation for each node by integrating both local and global contextual information. The local information is captured through a graphical convolution operation, which aims to obtain the immediate neighborhood characteristics of a node in the user graph. It focuses on local connections and features between a node and its neighbors to get node-level local contextual information, which can be defined as:

$$\mathbf{y}_i^{l+1} = \sigma\left(\sum_{j \in N_i} \frac{1}{w_{ij}} \mathbf{W}^l \mathbf{y}_j^l + b^l\right), \tag{3.4}$$

where $\mathbf{y}^l$ represents the hidden expression of node e_i at layer l, N_i summarizes all one-hop neighborhoods of node e_i, $\mathbf{W}^l$ and b^l are the weight matrix and bias vector respectively, and σ is the activation function. The term $1/w_{ij}$ indicates the normalized edge weight between nodes e_i and e_j, which can be obtained by learning from a data source.

For capturing global information, our AGGC model conducts the graph attention mechanism between each layer of the local graphical convolution operation. The global attention weights a_{ij}^l between nodes e_i and e_j at layer l are computed as: s

$$a_{ij}^l = \text{softmax}\left(\frac{\mathbf{y}_i^l \cdot \mathbf{y}_j^l}{\sqrt{d}}\right), \tag{3.5}$$

where $\cdot$ means the dot product operation. The attention weights determine the global importance of node connections and enable the model to focus on global contextual information.

At last, the updated hidden expression $\mathbf{y}_i^{l+1}$ at layer $l+1$ is learned by incorporating the local and global information using a linear transformation:

$$\mathbf{y}_i^{l+1} = \mathbf{W}_{local}^l \mathbf{y}_i^l + \mathbf{W}_{global}^l \sum_{j \in E} a_{ij}^l \mathbf{y}_j^l + b^l, \tag{3.6}$$

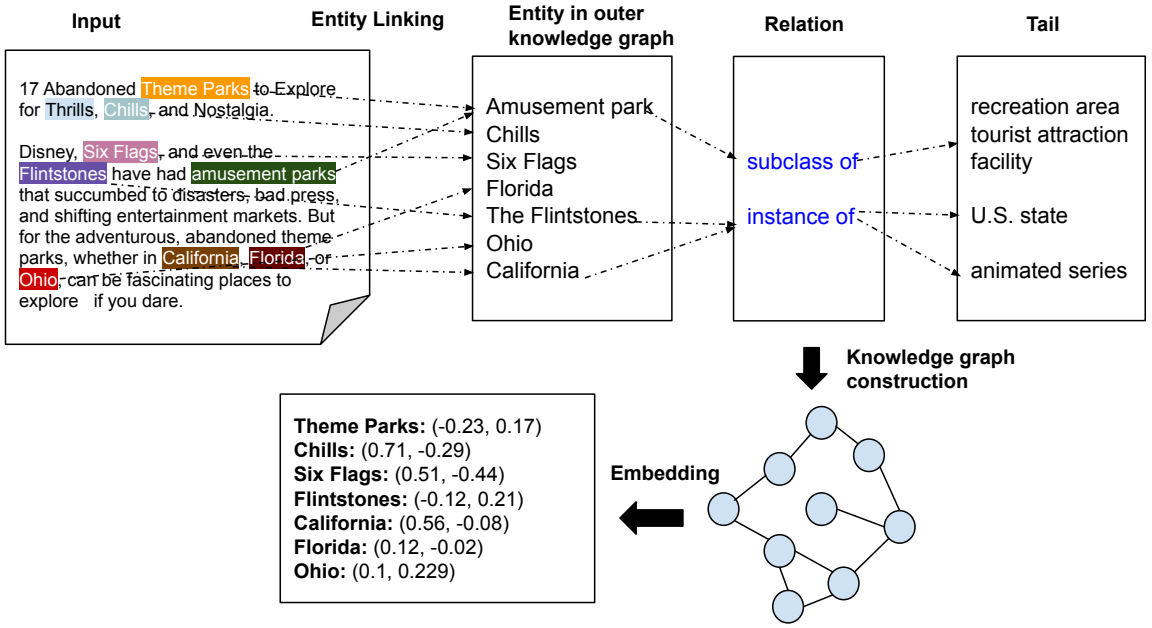


Fig. 3.6: An Example of User KG Construction and Representation

where $\mathbf{W}^l_{local}$, $\mathbf{W}^l_{global}$, and b^l refer to learnable parameters.

3.4.4 KG Distillation and Construction

In the proposed DOR, three approaches are applied to address the issue of inadequate entities and relations distillation and limited information delivery. First, a global KG is integrated with the existing data as input to the DOR model, addressing the cold-start problem, enhancing generalization capability, and increasing model robustness. Additionally, this global KG expands content coverage by filling distributional gaps in the original data, enhances diversity with a broader selection, and alleviates filter bubble effects by offering users access to content beyond excessive specialization. Second, word embedding techniques are employed to represent words in each input, capturing their low-order relations.

An example of the textual input KG extraction and construction process is illustrated in Figure 3.6. The input is a text composed of an input title or overview, with several keywords. Triples (subject, predicate, object) are extracted from the input. The inadequate triple extraction may lead to information sparsity in the experiment. I employ the entity alignment technique to link with corresponding entities and relations in the sub-graph (Wikidata KG). Based on these entities, their neighbors and relations are also explored. All the extracted information is then used to construct a KG of input.

3.4.5 Global-Observation Mechanism

The Global Observation Mechanism (GOM) in the DOR system receives the content and contextual features of the input from the local observation mechanism. The focus of this module is to study the different influences on the user's behavior from each piece of input. Therefore, the GOM is a DL model with a self-attention mechanism that analyses the mutual influence between users' beliefs and each piece of input. This module allows investigation of how users' perceptions and behaviors are affected by the input.

Specifically, the GOM takes into account the dynamic nature of users' beliefs. It incorporates a sequential characteristic that considers each historical reading record as a snapshot of the user's evolving belief network. The GOM receives the representations of the current input from the local observation mechanism and uses the user's current belief network to output the user's preferences. Thus, it models how the user's perception of current input is influenced by their previous reading history and belief network, which makes our model more adaptive to the user's dynamic interests.

Mathematically, given user i embedding U_{i_e} and the candidate input embedding N_{i_e}, the probability of user i clicking input N_i, i.e., p_{i,N_i}, is estimated through a general DNN $\mathbf{D}$:

$$p_{i,N_i} = \mathbf{D}(U_{i_e}, N_{i_e}) \tag{3.7}$$

3.5 Experiment and Analysis

In this section, I begin by assessing the performance of the proposed DOR model on three real-world datasets, comparing it with several established neural recommendation methods. Following this, an ablation study is conducted to validate the contribution of the dual observation mechanisms within the model. I then examine the influence of selected parameters on the model's overall performance. The DOR model exemplifies responsible recommendations by promoting content diversity, emphasizing its dedication to varied user experiences. Furthermore, an experiment was conducted to prove the responsible behavior of the DOR model. Finally, I discuss key insights derived from these experiments.

3.5.1 Experiment Setup

Dataset

This section describes the experimental setup for evaluating the performance of the proposed DOR. In the experiments, I leverage three real-world datasets from two sources: the Mi-

crosoft News datasets (MIND)[2] and IMDB dataset[3].

- **MIND Dataset** is a publicly available and large-scale news recommendation dataset obtained from Microsoft News. It includes two versions: small and large. The small version consists of 50,000 samples, while the large version contains 1 million users with 15 million impression logs. Each impression log represents a user's reading behavior and contains information such as user ID, timestamp, category, subcategory, title, abstract, and click behavior.

- **IMDB Dataset** consists of movie rating records derived from users' past behaviors. This dataset provides movie-related information, including movie ID, genres, titles, and overviews. Furthermore, it includes user ratings for target movies and their respective user IDs. In our experiments, I consider the rating degree of a movie as an indication of the user's interest. Ratings range from 0 to 5, where a rating of 5 represents a high level of interest, and a rating of 0 indicates no interest. I define a standard interest degree at 2.5, where ratings above this threshold imply user interest (clicked movie), while ratings below 2.5 suggest a lack of interest. By leveraging the IMDB dataset, I aim to enhance the recommendation capabilities of the DOR system by incorporating movie-related content and user preferences.

- **Datasets Statistics**. Table 3.1 provides an overview of the sizes and properties of three different datasets: IMDB, MIND-small, and MIND-large, showing the statistics related to the number of users, user behaviors, words, entities, and specific constraints on the data. The maximum number of words per title represents a constraint on the length of article titles or movie overviews. The titles or overviews in all three datasets have a maximum length of 20 words. Additionally, there are constraints on the maximum number of history entries and impression logs per user.

Experimental Settings

The proposed model and baselines are constructed by utilizing TensorFlow [1]. The experiments are configured with the intent to evaluate the effects of changing the number of epochs,

[2]https://msnews.github.io/

[3]https://www.kaggle.com/datasets/meastanmay/imdb-dataset?select=tmdb_5000_movies.csv

Dataset	IMDB	MIND-small	MIND-large
No. users	333	5,0000	711,222
No. behaviours	105,340	230,117	2,232,748
No. words	18,281	70,975	101,221
No. entities	1,843	57,023	372,124
Maximum No. words per title	20	20	20
Maximum No. history per user	50	50	50
Maximum No. impression log per user	100	100	100

Table 3.1: Statistics of datasets.

which ranged from 0 to 8, across all datasets. I adopt a batch size of 20 and establish a learning rate of 0.0001 for these trials.

To verify the contribution of the dual-observation (DO) mechanism, I give an experimental setup where the representation of the user's belief is removed from the DOR model. This serves as the local observation mechanism of the model. Additionally, I compare the performance of our proposed High-order Relation Model with other classical graph-based embedding models in the DOR.

In the experiments, I employ word embeddings of a 300-dimension specification. These embeddings are acquired through the GloVe2Vec method [28]. In addition to this, I evaluate the performance of other embedding techniques, including Word2Vec [28] and BERT [28].

The DO mechanism extracts contextual data and is trained using 16 layers. The context embedding dimension for each entity was consistently set to 300. The input of the DO module is pre-processed via the TransR model, having a dimensionality of 300, and is trained using a series of 10 batches. The DOR model is trained using the Adam optimization technique [52], with log loss optimization as the primary objective function.

3.5.2 Evaluation Metrics

In the experiments, I employ several widely used metrics commonly employed in the field of RSs [112].

- AUC (Area under the ROC Curve): AUC measures the probability that randomly chosen related items will rank higher than randomly chosen unrelated items. A higher AUC value indicates that the model can better distinguish between related and unrelated items.
- MRR (Mean Reciprocal Rank): MRR indicates the mean value of the reciprocal rankings of multiple query statements. It measures the effectiveness of a ranking system, with a higher MRR indicating a higher level of effectiveness.

- NDCG (Normalised Discounted Cumulative Gain): NDCG measures the ranking quality of a recommendation system. The principle of NDCG is that highly correlated products should rank higher than unrelated products. A higher NDCG value indicates a better ranking of related products.
- NDCG@5: The NDCG@5 metric calculates the DCG of the first five recommendations.
- NDCG@10: The NDCG@10 metric calculates the DCG of the first ten recommendations.

3.5.3 Baseline Methods

To evaluate the performance of our model, I compare it against several widely used baselines in the field of recommendation. These baselines include:

- **DKN** [210] represents a news recommendation system, employing attention networks to procure entity-word level representations. This model enhances the modeling of pertinent information by utilizing a dual-attention mechanism.
- **NRMS** [228] introduces an innovative approach for extracting user interests, in situations where user interest data is scarce. The proposed model in NRMS considers both the news title and abstract, thereby creating a comprehensive perception of users' reading interests.
- **NAML** [228] presents a novel neural news recommendation strategy, applying attentive multi-view learning to assimilate diverse types of news information into the representation of news.
- **LSTUR** [10] advocates a neural methodology for news recommendation that acknowledges immediate and lasting user interests. Utilizing the GRU (Gated Recurrent Unit), it captures short-term user preferences based on recent news engagement while considering long-term user interests.
- **FIM** [208] integrates multi-grained representation and matching methodologies to identify fine-grained interest signals through interactions among news articles at various semantic layers.
- **UNBERT** [250] addresses the cold-start issue in news recommendation by leveraging an out-domain data pre-trained model. It integrates multi-grained user-news matching

signals at both the word and news level via WLM (Word-Level Matching) and NLM (News-Level Matching) strategies, respectively.

- **MINER** [119] model utilizes a poly attention scheme to derive multiple user preference vectors, capturing various user interest facets through attention. It also implements a disagreement regularization technique to boost the diversity of the learned interest vectors. A category-aware attention weighting strategy is further adopted to adjust the significance of historical news based on category resemblance.

3.5.4 Performance Evaluation

In this section, I evaluate the proposed DOR model by comparing it against a number of state-of-the-art baselines. Table 3.2 demonstrates the performance of various models on three real-world datasets: MIND-Small, MIND-Large, and IMDB. As can be observed from this table, the DOR-Glove model achieved the highest performance in AUC, MRR, NDCG@5, and NDCG@10 across all datasets. "Improv. *min*" indicates the percentage by which the DOR-Glove model outperforms the MINER model in terms of expressiveness. "Improv. *max*" refers to the percentage by which the DOR-Glove model outperforms the DKN model regarding expressiveness.

	MIND-Small				**MIND-Large**				**IMDB**			
Model	**AUC**	**MRR**	**NDCG@5**	**NDCG@10**	**AUC**	**MRR**	**NDCG@5**	**NDCG@10**	**AUC**	**MRR**	**NDCG@5**	**NDCG@10**
DKN	0.629	0.2837	0.3099	0.3741	0.6407	0.3042	0.3292	0.3866	0.6784	0.2408	0.8407	0.5892
FIM	0.6502	0.3026	0.3291	0.3910	0.6787	0.3346	0.3653	0.4221	0.679	0.2383	0.8345	0.852
NRMS	0.6563	0.3096	0.3413	0.4052	0.6766	0.3325	0.3628	0.4198	0.6682	0.2356	0.8334	0.858
LSTUR	0.6587	0.378	0.3395	0.4015	0.6708	0.3236	0.3515	0.4039	0.6232	0.2488	0.8313	0.8396
NAML	0.6612	0.3153	0.3488	0.4109	0.6646	0.3275	0.3566	0.414	0.6774	0.2386	0.8336	0.8542
UNBERT	0.6762	0.3172	0.3475	0.4102	0.7068	0.3568	0.3913	0.4478	0.708	0.2595	0.8417	0.8724
MINER	0.6961	0.3397	0.3762	0.439	0.7151	0.3618	0.3972	0.4534	0.7123	0.2599	0.8611	0.8692
DOR-BERT	0.7262	0.3046	0.4469	0.5073	0.8586	0.3695	0.4508	0.6306	0.7202	0.2635	0.8615	0.90
DOR-Word2Vec	0.7576	0.3778	0.5296	0.5948	0.86	0.3729	0.5539	0.6516	0.7223	0.2689	0.8667	0.9054
DOR-Glove	**0.7917**	**0.3909**	**0.5539**	**0.6078**	**0.8701**	**0.4714**	**0.5731**	**0.6775**	**0.7309**	**0.2731**	**0.8688**	**0.9149**
$Improv._{min}$	0.1373	0.1507	0.4723	0.3845	0.2167	0.3029	0.4428	0.4942	0.0148	0.0364	0.0089	0.0525
$Improv._{max}$	0.2586	0.3778	0.7873	0.6246	0.3580	0.5496	0.7408	0.7524	0.0773	0.1341	0.0334	0.5527

Table 3.2: Performance comparisons.

To explain the results further, the DOR-Glove model obtained an AUC score of 0.7917 and 0.8701 on the MIND-Small and MIND-Large datasets, respectively, indicating its strong

predictive capability. Moreover, it achieved a high MRR of 0.4714, implying the successful ranking of relevant documents higher in the recommendation list. The NDCG@5 and NDCG@10 scores of 0.5731 and 0.6775 illustrate the model's effectiveness in promoting relevant documents to the top of the recommendation list.

On the other side, the DKN model demonstrated competitive performance on both MIND datasets, with AUC scores of 0.629 and 0.6407, respectively. However, it performed lower in MRR and NDCG metrics than the DOR-Glove model, implying that its ranking and relevance were not as strong. The FIM, NRMS, LSTUR, NAML, and UNBERT models achieved higher AUC scores than DKN, indicating superior discrimination ability. However, these models fell short in MRR and NDCG metrics, suggesting they may not rank relevant items as accurately as others. Interestingly, the DKN model outperformed the NRMS and LSTUR models on the IMDB dataset, attaining a 0.6784 AUC score. Among all baselines, the MINER model consistently delivered high performance across different metrics on all datasets.

As for DOR-BERT and DOR-Word2Vec, the variants of the DOR model, demonstrated competitive performance on the MIND-Small dataset, with higher AUC scores than other models. However, their performance in MRR and NDCG metrics was not as strong as the DOR-Glove model. In particular, the DOR-BERT model displayed weak performance on the MRR metric.

Based on the results of this experiment, the DOR-Glove model demonstrates superior performance than the baseline methods. Its strengths stem from effectively integrating dual observations when building user preference. It also captures global semantic relationships and contextual information within the text data, resulting in more precise and relevant recommendations.

3.5.5 Ablation Studies

In this section, I conducted three ablation experiments using MIND-Large and IMDB datasets to assess individual contributions of the various modules within the DOR model, enabling us to discern their specific impacts on performance and understand their importance in enhancing recommendation outcomes.

The first ablation experiment aims to evaluate the contribution of the proposed Dual Observations (DO) mechanism. The second ablation experiment examines the effectiveness of the proposed Attention-based Graph-enhanced Global Contextual (AGGC) model. The third ablation study investigates the influence of the classic translational distance models that can be adopted in DOR, including TransE, TransH, and TransR.

Ablation Study 1: By comparing the performance of the DOR model with and without DO, I aim to quantify the contribution of DO in our model. Recall that the DO mechanism employs a comprehensive two-observed pattern encompassing local and global perspectives. A model without the DO mechanism indicates this model will neglect the importance of the out-domain semantic information and the user's belief representations.

Table 3.3 demonstrates the results of the first ablation study, where MIND-Small, MIND-Large (represented as $MIND\uparrow$), and IMDB datasets are utilized.

Dataset	AUC	MRR	NDCG@5	NDCG@10
$MIND_{Non-DO}$	0.7522	0.3798	0.5341	0.5968
$MIND_{DO}$	**0.7917**	**0.3909**	**0.5539**	**0.6078**
$MIND\uparrow_{Non-DO}$	0.8384	0.2903	0.5688	0.6566
$MIND\uparrow_{DO}$	**0.8701**	**0.4714**	**0.5731**	**0.6775**
$IMDB_{Non-DO}$	0.6354	0.2331	0.8539	0.8390
$IMDB_{DO}$	**0.7309**	**0.2731**	**0.8688**	**0.9149**

Table 3.3: Ablation study on dual observation mechanism

As can be seen from Table 3.3, for the MIND dataset, the results explicitly show that the model with the DO mechanism $(MIND_{DO})$ achieved a higher AUC (0.7917) compared to the model without DO $(MIND_{Non-DO})$ (0.7522). Similarly, the DO model outperformed the Non-DO model in terms of MRR, NDCG@5, and NDCG@10. In the case of the MIND-Large dataset, the mode with DO demonstrates even better performance, with a higher AUC (0.8701), MRR (0.4714), NDCG@5 (0.5731), and NDCG@10 (0.6775) compared to the that of Non-DO. Lastly, in the IMDB dataset, the model with DO achieved an AUC of 0.7309, MRR of 0.2731, NDCG@5 of 0.8688, and NDCG@10 of 0.9149, while it had poor performance without DO. This further highlights the effectiveness of the DO mechanism, particularly when applied to larger datasets.

The results from this ablation study consistently demonstrate that incorporating the DO mechanism leads to improved performance across various evaluation metrics. The higher AUC, MRR, and NDCG scores obtained by the DO models indicate their superior effectiveness in capturing and utilizing dual observations compared to those without the DO mechanism.

Ablation Study 2: I proposed the Attention-based Graph-enhanced Global Contextual (AGGC) model to capture high-order relation representations. In this ablation study, I replaced the AGGC model with three alternative graph embedding models, namely GCN, GAT, and GraphSAGE, to evaluate their performance within our model.

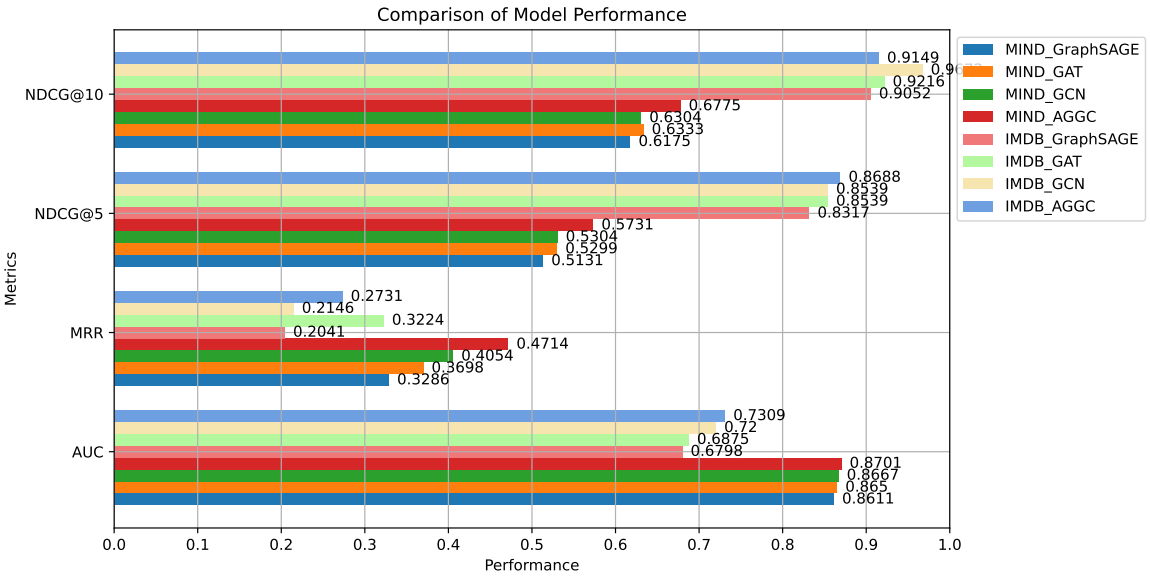


Fig. 3.7: Ablation study on diverse graph representation models

Fig. 3.7 illustrates the performance of different graph embedding models on various metrics, namely AUC, MRR, NDCG@5, and NDCG@10. The y-axis represents the metrics, while the x-axis represents the performance values ranging from 0 to 1. Each line in the graph corresponds to a specific model and is assigned a specific color.

The models MIND_GCN, MIND_GAT, and MIND_AGGC exhibit similar and higher AUC scores compared to MIND_GraphSAGE. Among the IMDB models, IMDB_GCN and IMDB_GAT perform better than IMDB_GraphSAGE, while IMDB_AGGC demonstrates the highest AUC score. MIND_AGGC consistently achieves the highest MRR score among all models. MIND_GCN and MIND_GAT also perform relatively well. However, for the IMDB dataset, IMDB_AGGC does not perform as well as AGGC used on the MIND dataset. MIND_AGGC shows the highest NDCG@5 and NDCG@10 scores, indicating its superior performance. MIND_GCN and MIND_GAT also exhibit competitive performance. IMDB_AGGC achieves the highest NDCG@5 and NDCG@10 scores among the IMDB models, followed by IMDB_GAT.

In conclusion, the graph reveals that MIND_AGGC consistently outperforms the other models across all metrics. It showcases the effectiveness of the AGGC model architecture in the MIND dataset. Similarly, among the IMDB models, IMDB_AGGC and IMDB_GAT demonstrate relatively better performance compared to IMDB_GraphSAGE and IMDB_GCN.

Ablation Study 3: I analyze the impact of incorporating these different translational distance models on the overall performance of the DOR model.

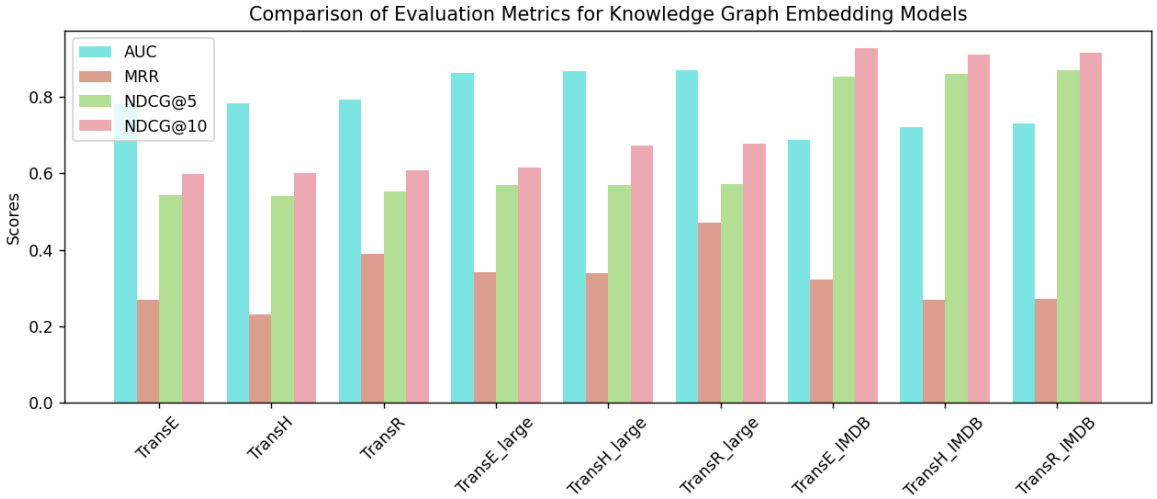


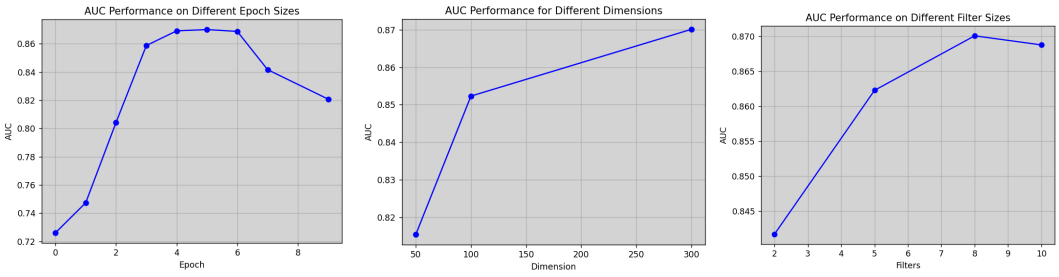
Fig. 3.8: Ablation study on different translational embedding models.

This graph's "large" variate indicates the MIND-Large dataset, and 'TransE_IMDB' expresses that the IMDB dataset trains the transE-based DOR model. Figure 3.8 presents a comparative analysis of various KG embedding models, including TransE, TransH, and TransR, when applied to DOR. The evaluation metrics, AUC, MRR, NDCG@5, and NDCG@10, are plotted on the y-axis, while the x-axis signifies the different models.

The results reveal the performance differences between the models across the four evaluation metrics. TransE_large, TransH_large, and TransR_large show a dominant performance over the rest. The "large" suffix in these models signifies their training on a larger-scale dataset, specifically the MIND-Large.

TransR leads in performance, consistently achieving peak scores in AUC, MRR, NDCG@5, and NDCG@10. This suggests that TransR excels at modeling intricate relationships and encapsulating the semantic interactions between entities and relations in the DOR. TransE also demonstrates competitive performance, with high scores in all metrics. While its scores fall slightly short compared to TransR and TransH, it maintains a robust overall performance. TransH's performance mirrors that of TransE, indicating its effective capture of the interactions between entities and relations when applied to DOR.

TransR shows superiority across all evaluation metrics. Thus, in the context of DOR, TransR is regarded as a promising choice for KG embedding to address complex relationships and semantic interactions.



(a) AUC score on different epoch sizes (b) AUC score for varying dimension numbers (c) AUC score for diverse filter sizes

Fig. 3.9: Parameter analysis on AUC scores

3.5.6 Parameter Analysis

In this section, I aim to analyze the impact of various parameters on our proposed DOR model. Specifically, I investigate the effects of epoch numbers, filter sizes, and dimension numbers on the performance of the DOR model using the MIND-Large dataset.

First, I conduct experiments with different epoch sizes from 0 to 8 to evaluate the performance of the DOR model. This allows us to examine the model's behavior and performance under varying training durations. Second, I explore the influence of filter sizes in the DOR system. Specifically, I experiment with 5, 8, and 10 filters, respectively, as the number of filters for each size in the DOR system. This analysis helps us understand the impact of filter sizes on the model's ability to capture relevant information and make accurate recommendations. Third, I examine the effects of dimension numbers on the performance of the DOR system. I vary the dimensions for all features of the DOR model, specifically exploring dimensions 50, 100, and 300. By assessing the model's performance with different dimension settings, I am capable of obtaining an optimal dimensionality for feature representations.

Figure 3.9a demonstrates the model's performance over multiple epochs, with the x-axis representing the epochs and the y-axis representing the AUC values. The AUC value commences at 0.7261 (Epoch 0) and displays an upward trend through training, peaking at 0.8701 (Epoch 5). Subsequent epochs exhibit slight fluctuations in the AUC value, suggesting that the model's performance stabilizes and may not see significant improvements beyond Epoch 5.

In a second parameter analysis, the dimensions are varied according to Figure 3.9b, which reveals that AUC performance differs based on the dimensions set. There's an upward trend in AUC values as the dimensionality increases: 0.8154 for dimension 50 (moderate

performance), 0.8523 for dimension 100, and 0.8701 for dimension 300, indicating an improved model performance with increased dimensionality.

Lastly, an analysis of filter sizes (2, 5, 8, and 10) is conducted to evaluate the performance of the DOR model. Figure 3.9c reveals a gradual increase in AUC performance corresponding to the number of filters, reaching its peak at 8 filters with an AUC value of 0.8701. Although the AUC slightly decreases to 0.8688 at 10 filters, it remains close to the highest value, suggesting that increasing the number of filters within a certain range may have little impact on performance.

These three parameter analyses reveal the interplay between various parameters and the model's performance, providing critical insights for optimizing the proposed DOR.

In summary, our experiments demonstrate that the DOR model holds a clear advantage in recommendations when utilizing the dual-observation mechanism. Our results confirm that it is essential to fully observe the features of items. Additionally, our experiments demonstrate the importance of considering the mutual influence between the user's belief network and the items in the recommendation process.

3.5.7 Impact of High-Order Relations Model: Global Knowledge-Enhanced Data

In this experiment, the role of HRM within the DOR model is investigated by incorporating global external data into the HRM component. By integrating external information from sources such as Wikidata, the DOR model aims to expand the embedding space coverage, thereby enhancing content diversity and mitigating filter bubble effects. To visualize and analyze the influence of global knowledge-enhanced data on entity distribution, principal component analysis (PCA) [82] was applied to reduce the high-dimensional entity embeddings generated by the AGGC model. This dimensionality reduction allows for a comparative visual representation of the original dataset (Original Data) and the global knowledge-enhanced dataset (Wiki Data), enabling the observation of how high-order relations impact the overall responsible distribution.

Figure 3.10 illustrates the embedding distributions for both the Original Data (left) and the Global Knowledge-enhanced Data (right). In each plot, each point represents an entity within the respective dataset, with positions determined by their reduced-dimensional coordinates. The relative distances between points reflect semantic similarity, entities positioned closely together are more similar in the high-dimensional embedding space, while those farther apart indicate greater dissimilarity.

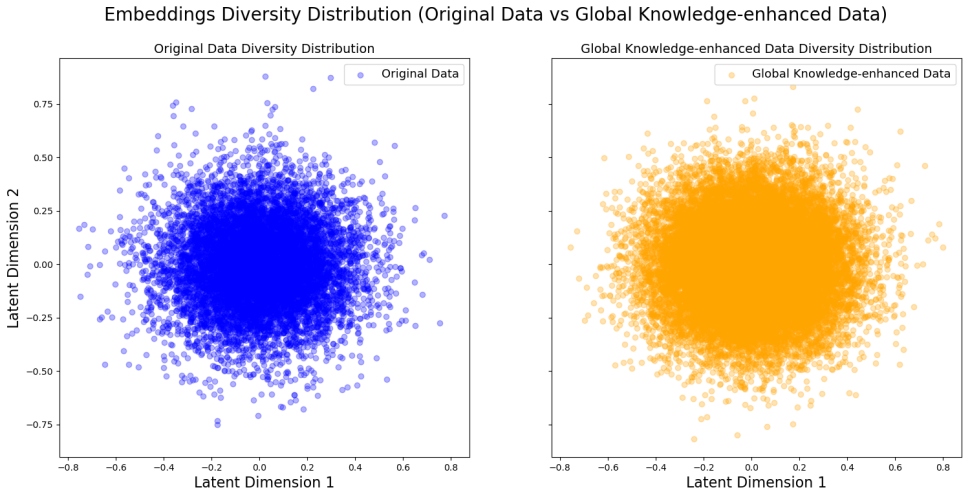


Fig. 3.10: Entities Diversity Distribution

The visual comparison between the two projections reveals the impact of high-order relations introduced through global knowledge-enhanced data. In the Original Data projection (left), entities are more tightly clustered, indicating limited diversity and potential content gaps. In contrast, the Global Knowledge-enhanced Data projection (right) displays a broader and more even distribution across the embedding space, effectively filling areas that were sparse in the Original Data. This expanded distribution demonstrates that the HRM component, enriched by high-order relations external data, introduces a wider range of entities and connections inferred beyond the explicit content of the articles.

This distribution pattern supports the notion that high-order relations facilitated by global knowledge-enhanced data play a crucial role in reducing filter bubbles by exposing users to a broader array of topics. The more dispersed distribution achieved with high-order relations enables the RS to present a balanced mix of familiar and novel content, thereby reducing the risk of over-specification and promoting a more comprehensive and diverse recommendation experience.

This experiment demonstrates that incorporating global knowledge-enhanced high-order relations into the DOR model's HRM component significantly improves content diversity and mitigates filter bubble effects. The global knowledge-enhanced data not only broadens the range of entity connections but also enhances the RS's capacity to deliver varied content, fostering a more responsible recommendation environment.

3.5.8 Discussion

Among evaluations, three real-world datasets were adopted to compare the performance of the DOR model against well-known neural recommendation methods. This comparison provides a comprehensive assessment of the model's effectiveness. The results showed that the DOR model outperformed the competitors, indicating its superiority in involving dual observations, capturing global semantic relationships, and utilizing contextual information for more accurate and relevant recommendations.

Based on the experiment results, it can be observed that the proposed DOR model is superior to other baselines. This reflects that the DOR model can generate more accurate results and provide more relevant recommendations. This has vital implications for solving information overload in RSs and improving user satisfaction. These experimental results show that by introducing a dual-observation mechanism, i.e., a local observation mechanism that fully considers contextual semantic information and hybrid-domain features and a global observation mechanism that adequately captures the mutual propagation between the user's belief network and outside data sources, our DOR model can better dig into vital user preferences. This is critical for personalizing recommendations and improving user experience.

Furthermore, my experiments demonstrate the effectiveness of high-order relation representation in the DOR model. By employing graph representation models, especially our AGGC model, We can depict higher-order relational representations. This is critical for RSs, as it can explore more complex semantic information among items, thus providing more accurate, relevant, and personalized recommendations. Additionally, the flexible use of the global observation mechanism in the model can further reflect the interaction between the user's existing KG and external information sources.

One of the key findings from this study is the positive impact of Wiki-enhanced high-order relations on the DOR model's diversity and responsibility in recommendations. By integrating external data from Wikidata into the HRM component, we observed a significant improvement in the diversity of entity embeddings. The Wiki-enhanced data expanded the embedding space, which is evident in the visual comparison between Original Data and Wiki Data projections. The broader and more even distribution achieved with Wiki-enhanced data demonstrated that high-order relations could fill gaps in the original dataset, providing a more comprehensive and diverse content recommendation environment. This enrichment allows the DOR model to introduce new and varied content, thereby mitigating the risk of filter bubbles and reducing over-specification.

This distributional expansion facilitated by Wiki-enhanced data has important implications for responsible recommendations. By exposing users to a wider array of topics and

perspectives, the model can deliver a balanced mix of familiar and novel content, which supports a more responsible RS. This approach not only improves user satisfaction by preventing content saturation but also fosters a more dynamic and exploratory user experience, encouraging users to discover content outside their established preferences.

These findings have important implications for advancing domain knowledge and RSs. My experimental results illustrate that the performance of RSs can be significantly improved by incorporating the dual-observation mechanism, particularly with the support of high-order relations through external data sources. By enhancing both the accuracy and diversity of recommendations, our DOR model demonstrates a potential pathway toward more responsible and user-centered RSs that mitigate common issues such as information overload and filter bubbles.

3.6 Conclusion and Future Work

In this research, I proposed a new recommendation method called the DOR model, which uses a dual observation mechanism to tackle challenges like cold start, data sparsity, and over-specialization. Extensive experiments on real-world datasets, especially for news recommendations, have shown the model's effectiveness in improving recommendation accuracy compared to leading methods. By combining both local and global observations, the DOR model strengthens the connections within user confidence networks and integrates outside information, providing a more complete view of user preferences. However, while the DOR model effectively improves accuracy, it also brings to light the need to balance personalized recommendations with diversity to offer users a broader range of content.

As this research progresses, the focus will shift from maximizing accuracy and personalization to addressing the social impacts of RSs, particularly the issue of filter bubbles. Filter bubbles occur when users are repeatedly shown similar information, which reinforces their existing beliefs and may lead to biased views. This effect poses risks such as social division, the spread of misinformation, and a decline in critical thinking. Therefore, it is essential to find methods that reduce these effects, ensuring that RSs not only match user preferences but also present diverse information.

To further tackle these challenges, the next chapter introduces a graph-based approach designed to reduce filter bubbles and biased exposure in RSs. By using advanced techniques in graph modeling, this approach aims to create a flow of information that exposes users to a wider range of perspectives while still respecting their interests. This shift reflects the growing importance of building RSs that are not only accurate but also socially responsible, providing users with a well-rounded experience in today's complex information environment.

Chapter 4

A Graph-Based Responsible Approach to Reducing Filter Bubble Effects in Recommendations

4.1 Introduction

Recommendation systems play an important role in shaping users' access to diverse information on the Internet [111]. However, most of the existing preference-based recommendation systems continuously suggest items that are similar to users' previous experiences, leading them to a homogeneous information environment. This is known as the "filter bubble" phenomenon [137]. Prolonged exposure to such "filter bubbles" can result in the development of extreme and imbalanced beliefs, hindering the formation of a comprehensive understanding and amplifying ideological biases [65]. In the context of responsible artificial intelligence (AI), it becomes crucial to develop AI recommendation systems that provide a diversity of content and viewpoints, rather than reinforcing existing user preferences, to mitigate "filter bubbles" [67].

Existing research works on mitigating the filter bubble in recommendation systems can be grouped into two main strategies: algorithm-focused and human-focused approaches [6]. Algorithm-focused strategies advocate for promoting content diversity at both the in-processing and post-processing stages [6]. These methods leverage diverse techniques, including explanation-based diversity recommendation [245], community-aware models [83], category-based diversification algorithms [132], the Diversified GNN-based Recommendation system (DGRec) [242], and graph-based user-item interaction methods [126]. Notably, graph-based approaches, relying on user preferences and category diversity insights, aim to

enhance recommendation quality [85]. However, these strategies, particularly those driven by algorithms, may unintentionally ignore the essential role of human decision-making processes. In contrast, human-focused strategies emphasize individuals [6]. Techniques such as nudging-based recommendations [99, 103] indirectly influence user decisions and behaviors. Despite focusing on users, these models often encounter challenges in effectively broadening users' interests beyond their preferred topics. Unlike existing mitigation approaches, this chapter integrates the strengths of both algorithmic and human-focused strategies and presents the Responsible Graph-based Recommendation framework, namely RGRec, to mitigate "filter bubbles" by mildly moderating users' extreme beliefs, thus exposing them to a more diverse scope of information.

RGRec is a graph-based approach that serves as an intermediary between recommendation systems and users. Its primary objective is to effectively address the "filter bubbles" issues by bridging the gap between user preferences and the delivery of diverse content recommendations. The system comprises three key modules: the Multi-faceted Reasoning-based "filter bubbles" Detection module (FBDetect), the Belief Nudging module, and the Generative Artificial Intelligence-based Recommendation Strategy Generation module (RecomGen). In FBDetect, a user's belief is represented as a heterogeneous graph [123] known as a belief network. FBDetect identifies users affected by "filter bubbles" by evaluating the balance between a user's belief toward a specific topic of information and the recommendations received from the system. If this balance is significantly skewed, the user is flagged as being impacted by "filter bubbles". The Belief Nudging module collaborates with users' belief networks to explore paths between topics that users favor and those they display less interest in. These explored paths serve as prompts for RecomGen to generate items for a nudging recommendation strategy. This strategy aims to gently introduce users to content they may have shown less preference for, fostering a more balanced exposure to diverse content. The collaboration among these three modules is iterative, continuously optimizing and adjusting to mitigate "filter bubbles".

The ultimate goal of this research is to gradually introduce users to a more diverse range of content, fostering belief harmony and enhancing the overall user experience. RGRec's innovative approach addresses the challenges of "filter bubbles" by proactively diversifying recommendations and promoting a more nuanced interaction between users and content.

- Firstly, I introduce a novel responsible approach, i.e., RGRec, designed to address the moderation of users' extreme beliefs and the mitigation of "filter bubbles" resulting from conventional recommendation approaches. To the best of our knowledge, RGRec stands out as one of the pioneering responsible recommendation methods explicitly focused on alleviating "filter bubbles".

- Secondly, I present the Multi-faceted Reasoning-based "filter bubbles" Detection module (FBDetect), a pivotal component within RGRec. FBDetect identifies users affected by "filter bubbles" and scrutinizes recommendation systems relying solely on user preferences. Our approach employs diverse methodologies to comprehensively analyse "filter bubbles", examining their existence and effects from various perspectives.
- Thirdly, I leverage the efficacy of nudging techniques to guide users in broadening their interests and promoting belief harmony. Our nudging process aligns with principles of libertarian paternalism, transparency, and democracy, thereby enhancing users' understanding of recommendations.
- Finally, I present the Generative Artificial Intelligence-based Recommendation Strategy Generation module (RecomGen) for crafting recommendation strategies aimed at mitigating "filter bubbles". This method leverages advanced graph-based techniques to learn and analyze user beliefs, systematically exploring potential paths to alleviate users' extreme beliefs by introducing a more extensive range of information and enhancing content diversity.

The rest of this chapter is organized as follows. Section 4.2 offers a comprehensive review of related literature, delving into recommendation systems, user belief bias, "filter bubbles", and nudging techniques. Section 4.3 elucidates essential definitions, notations, and concepts integral to our discourse. The methodology and framework underpinning our research are expounded upon in Section 4.4. Section 4.5 is dedicated to the elucidation of our experimental setup, with subsequent presentation and analysis of results. In Section 4.6, the findings are examined and discussed. Section 4.7 concludes the paper and outlines potential directions for future research.

4.2 Related Works

Personalized recommendation systems have been criticized as inadvertently creating "filter bubbles" [153], which constrain users' exposure to various perspectives and information, thus potentially leading to belief biases and societal fragmentation [132]. To mitigate this concern, many researchers and practitioners have focused on dismantling "filter bubbles", fostering diversity and democracy in recommendation systems, and facilitating users' belief harmony.

In this section, I review the relevant research works, deliberate on the filter bubble issue, explore the diversification of recommendation systems, and examine the prior research on nudge recommendations. Additionally, I will highlight the contributions of this study.

4.2.1 Filter Bubbles

Preference-based Recommendation Systems

Conventional recommendation systems prioritize the generalization of user preference, implying that these systems often recommend items to users based on their specific preferences and behaviors [84]. Techniques such as Collaborative Filtering (CF) [204], Content-Based filtering (CBF) [162], rule-based methods [236], or hybrid models [2] are commonly employed to analyze users' preferences and past behaviors. The recommendation system then suggests content that aligns closely with user preferences to enhance user satisfaction and engagement. However, this approach based on user preference may exacerbate the filter bubble issue, leading to ideological isolation and user bias. For example, Bryant et al. demonstrate that the YouTube algorithm, representative of a preference recommendation algorithm, exhibits a marked bias towards right-leaning political videos, including those espousing racist views propagated by the alt-right community [29]. Thus, it is important to address the limitations of current preference recommendations, boost the diversity of suggestions, and harmonize users' beliefs.

Mitigating Filter Bubble Effects

"Filter bubbles" emphasize the constraints of preference recommendation algorithms [110]. Dahlgren introduced the term "internet filters" to represent the phenomenon of "filter bubbles", which can have various negative effects on users, including a narrowed focus on personal interests, substantial reinforcement of confirmation bias, reduced curiosity, decreased exposure to diverse ideas and people, compromised understanding of the world, and a skewed perception of reality [49].

Addressing the negative effects of "filter bubbles" proves to be challenging, especially when considering the notable aspect of algorithmic bias. Chen et al. argue that the emergence of recommendation algorithm bias amplifies the experimental nature of user behavior data as opposed to observational [39]. Additionally, Dahlgren examines the recommendation algorithm bias and broadens the concept of bias into two facets, one originating from the recommendation algorithm and the other from users' behaviors [49]. Aside from algorithmic bias, another challenge in mitigating "filter bubbles" lies in their elusive nature [132]. Users often find themselves unaware of the filter bubble effect, which creates a homogenized view of the world. Specifically, they may not realize that their perspective differs from others in similar circumstances.

The growing influence of "filter bubbles" has raised increased concerns among researchers. A well-crafted recommendation system usually offers high accuracy while promot-

ing diversity; systems oriented solely towards accuracy may inevitably lead to filter bubble effects [242]. Contemporary research proposes several strategies for breaking "filter bubbles" by enhancing the diversity of recommendations. The research addressing "filter bubbles" with graphs is reviewed as follows.

Research in the scope of the graph. While the above-mentioned research has significantly contributed to alleviating "filter bubbles" and enhancing recommendation diversity, many researchers also advocate for the critical role of graph-based recommendation algorithms. These algorithms mitigate data sparsity and cold start issues and add an essential interpretability factor to recommendation systems [85]. Yang et al. introduce the Diversified GNN-based Recommendation System (DGRec), a graph-based recommendation system built on GNN, augmenting the diversity of recommended lists by improving the embedding generation process [242]. Tang et al. propose a temporal graph-based method to learn user evolving preferences in dynamic recommendation scenarios [242]. Additionally, Li et al. adopt a graph-based methodology by constructing a user-item interaction graph for data analysis to examine the existence of a centralized recommendation phenomenon [126].

In contrast, our model surpasses traditional methods by generating more diverse items instead of marginally varied ones, based on user preferences for diversity. I prioritize incrementally stimulating users' interest in items they may initially disregard without altering existing recommendation system algorithms. The proposed novel approach aims to counteract the filter bubble effect by considering user interest and disinterest beliefs, i.e., an aspect that has received minimal attention from researchers.

Detection of Belief Bias. Belief bias in reasoning refers to individuals' tendency to favor conclusions that align with their pre-existing beliefs [31]. This phenomenon is connected with the formation of online "filter bubbles", in which users tend to accept information that confirms their viewpoints and interests while rejecting alternative perspectives that challenge their beliefs [164, 92].

Existing methods proposed for belief bias detection include Information Source Diversity Analysis (ISDA) [132], User Interaction Pattern Analysis (UIPA) [218], Reinforcement Learning Methods (RLM) [126], and Social Network Analysis (SNA) [218, 44]. Considering the limited interpretability of RLM and the focus of SNA on alleviating echo-chamber effects rather than "filter bubbles", our research concentrates on investigating selective belief bias algorithms based on ISDA and UIPA. ISDA includes various detection metrics such as topology metrics and homophily metrics [132]. Likewise, UIPA includes several established detection metrics, including the coverage algorithm and the Majority Category Domination (MCD) algorithm [218]. Drawing inspiration from these metrics, I propose the FBDetect

model for dual verification of the authenticity of the filter bubble phenomenon, having the concept of "Entropy" [238] included to substantiate the existence of "filter bubbles".

4.2.2 Nudge Techniques and Responsible Recommendations

A *nudge* is a non-coercive intervention designed to influence behavior by modifying the context in which choices are made [27]. Such an intervention is usually transparent, optional, and responsible, enabling individuals to understand their choice consequences better and boost the likelihood of beneficial decision-making [24]. The core idea behind a nudge is to exploit individuals' beliefs and behavioral biases through various design strategies, such as providing incentives [24] and utilizing social influence [124, 205], directing them towards more favorable outcomes without constraining their freedom of choice [27, 232].

Recent research in recommendation systems has begun to explore the role of nudges. However, the majority introduces nudging recommendations from an AI-deprived perspective, implying a substantial absence or lack of AI technology in their research context. For example, Jesse et al. consolidate 87 nudging mechanisms at this AI-deprived level, including alterations in font size, the reputation of the messenger, and the visibility of information [99]. Joachim et al. propose a platform empowered by AI designed to nudge, influence, and guide the behavior of individuals with diabetes [103]. Furthermore, Sitar et al. propose an automated recommendation system. This system integrates managers' priorities and user feedback and utilizes graph structures to organize items based on descending order of priority, known as nudge concepts [184].

The recommendation systems mentioned previously have revealed the importance of establishing a responsible, graph-based nudging recommendation system. However, existing models are developed solely on user preferences, neglecting the influence of "filter bubbles". Unlike the existing preference-based approaches, this research aims to gently present more potential interests to users whom they may have yet to be genuinely interested in initially. Guiding user perceptions from one end of the graph (items users are highly interested in) to the other end (items users are less interested in), moderating user extreme beliefs, reducing user bias, and breaking the filter bubble effect, thereby allowing users to access a more diverse range of information.

4.3 Preliminaries

In this section, I introduce definitions, notations, and concepts used in this chapter. Key notations are listed in Table 4.1.

A recommendation environment is defined as $S = (U, A, C)$, where $U = \{u_1, ..., u_n\}$ represents a set of users, A refers to an AI-based algorithm for recommending items to users, and $C = \{c_1, ..., c_x\}$ signifies a set of pre-defined topics. Meanwhile, each c_x is associated with a set of aspects $C_x^{sub} = \{c_1^x, ..., c_k^x\}$, and each c_k^x is associated with an item set $I_k^x = \{i_1, ..., i_m\}$. An item can represent a news article or a movie description in real-world applications. To simplify the problem, in this chapter, I assume each item belongs to only one aspect, and each aspect is related to a single topic.

Definition 1. ***The belief network*** *of a user u_i is represented as a directed graph $G_i = (V_i, E_i)$. Specifically, $V_i = \{u_i\} \cup \hat{C}_i \cup \hat{C}_i^{sub}$ represents a set consisting of three distinct types of nodes, where $\hat{C}_i \subset C$ denotes a set of* ***topics*** *that u_i has engaged with, and $\hat{C}_i^{sub} = \{c_k^x | c_k^x \in C_x^{sub}, c_x \in \hat{C}_i\}$ represents different* ***aspects*** *related to $\hat{C}_i$. Meanwhile, $E_i = E_i^b \cup E_i^c$ represents a composite set of edges, where $E_i^b = \{e_{ix} | c_x \in \hat{C}_i\}$ comprises edges connecting from the user to each topic, and $E_i^c = \{e_{xk} | c_x \in \hat{C}_i, c_k^x \in \hat{C}_i^{sub} \cap C_x^{sub}\}$ consists of edges connecting from each topic to each of its aspects. In G_i, e_{ix} denotes u_i prefers a topic c_x, and the weight b_{ix} associated on e_{ix} represents the extent of u_i's belief towards a topic c_x. While e_{xk} indicates the affiliation relationship between each pair of topic and topic aspect, and its weight r_{ik}^x reflects the probability that u_i selects an item whose aspect is $c_k^x \in \hat{C}_i^{sub}$.*

The click probability r_{ik}^x is calculated using Equation 4.1, where $|\hat{I}_k^x|$ denotes the number of u_i interacted items whose aspects are c_k^x. In this case, given a belief graph G_i, the sum of all r_{ik}^x is 1.

$$r_{ik}^x = \frac{|\hat{I}_k^x|}{\sum_{c_{x'} \in \hat{C}_i} \sum_{c_{k'}^{x'} \in \hat{C}_{x'}^{sub}} |\hat{I}_{k'}^{x'}|} \tag{4.1}$$

After calculating all click probability r_{ik}^x, the belief degree b_{ix} of user u_i towards a topic c_x can be formulated as follows:

$$b_{ix} = -\sum_{c_k^x \in \hat{C}_i^{sub} \cap C_x^{sub}} r_{ix}^k \log_2(r_{ix}^k), \tag{4.2}$$

Definition 2. ***Topic similarity*** *$\rho(c_x, c_y)$ represents the similarity between two topics c_x and c_y, with symmetry $\rho(c_x, c_y) = \rho(c_y, c_x)$. The similarity measure ρ is defined within the range $[-1, 1]$, where a higher value of ρ indicates greater similarity between the topics.*

As each item is only associated with one topic, I can obtain topic embedding by aggregating related item embeddings. In this chapter, I obtain the topic embedding $\mathbf{c}_x$ by adopting the Hadamard product [138] to combine all corresponding item embeddings. The similarity between topic c_x and c_y is subsequently calculated by the cosine similarity [20]:

$$\rho(c_x, c_y) = \frac{\mathbf{c}_x \cdot \mathbf{c}_y}{\|\mathbf{c}_x\| \left\|\mathbf{c}_y\right\|}, \tag{4.3}$$

where $\mathbf{c}_x$ and $\mathbf{c}_y$ denote topic embeddings, $\mathbf{c}_x \cdot \mathbf{c}_y$ refers to the dot product of the topic embeddings, and $\|\mathbf{c}_x\|$ and $\left\|\mathbf{c}_y\right\|$ denote the corresponding Euclidean norms.

The primary emphasis of this research does not center around language embedding. In this chapter, our approach involves the utilization of a pre-trained language model [108] for the purpose of item embedding.

Definition 3. ***A recommendation prompt path*** *$p_{i,t}$ is a sequence of topics explored by a filter bubble-affected user u_i at a specific time step t, bridging the gap between topics $c_x^{SP_i}$ (u_i strongly preferred) and $c_y^{LP_i}$ (u_i less preferred) by introducing additional interacted topics. $p_{i,t}(k)$ refers to the k^{th} topic in a recommendation prompt path $p_{i,t}$. Furthermore, the topics within $p_{i,t}$ can be the keywords of prompts for the RecomGen module to generate contextually rich items for recommendation to u_i (see Definition 4).*

Definition 4. ***A recommendation list*** *$feed = \{feed_{original}, GI\}$ is a compilation of items recommended to users, where $feed_{original}$ is comprised of items suggested by the existing preference-based recommendation system, and GI includes contextually rich items generated by RGRec, based on an explored recommendation prompt path (refer to Definition 3). Given the size of $feed$, I introduce a weight parameter w to control the proportion of GI within the $feed$, balancing the mix of original and RGRec recommendations. $\check{M}_{i,t}$ signifies the set of items in a $feed$ accepted by user u_i from the initial time step to time step t. In contrast, $\bar{P}_{i,t}$ refers to the sequence of recommendation prompt paths shown to u_i, where the resulting items have been declined by the user over the same period.*

Incorporating *GI* in a $feed$ bridges the gap between filter bubble-affected users' preferred and less preferred topics. It introduces more interconnected topic-related items, moderates extreme beliefs, and encourages belief harmony.

4.4 The Framework of Responsible Graph-based Recommendation

Responsible Graph-based Recommendation (RGRec) is designed to guide users gently from a state of information imbalance to one of belief harmony. RGRec stands out for its incorporation of the "nudge" concept and the utilization of GAI to produce contextually rich items. This approach encourages users to explore interests in topics they may have originally shown less interest in, providing diverse options. Figure 4.1 illustrates the overview of RGRec's working process, using a practical example involving the user "U21538".

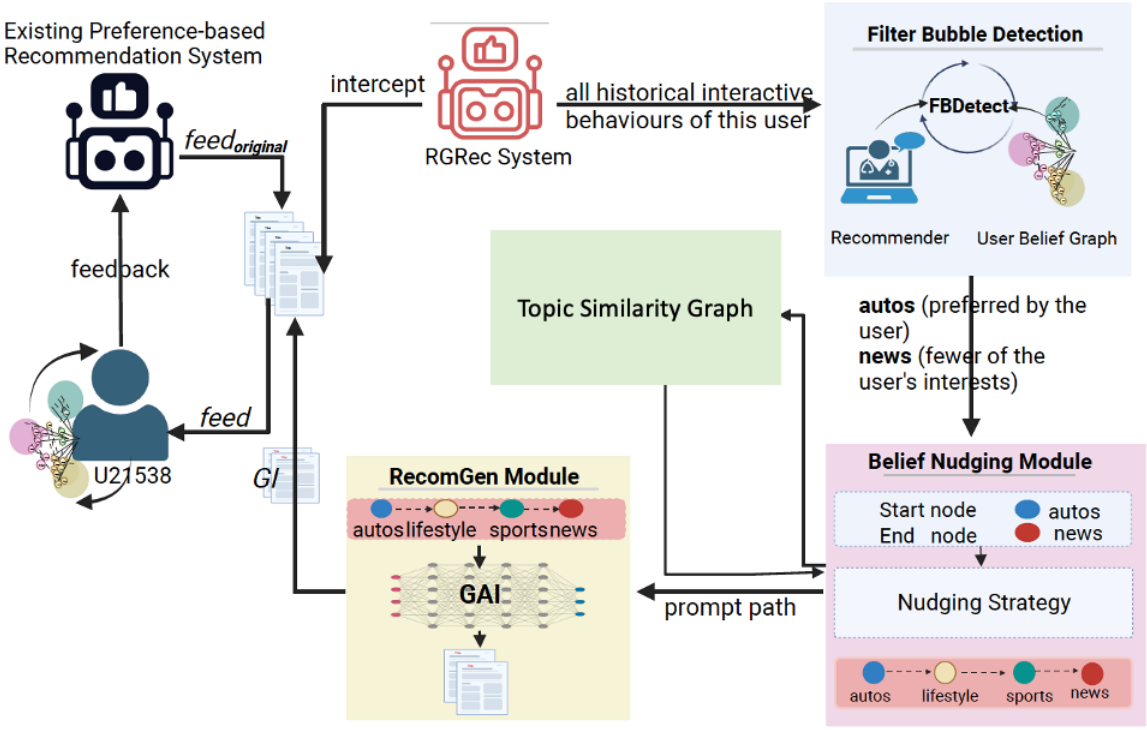


Fig. 4.1: Overall working process of RGRec for user "U21538".

As shown in Fig. 4.1, RGRec operates as a dynamic and interactive mediator between existing preference-based recommendation systems and the users, e.g., "U21538" in this example. It conforms to the principle of non-coercion, ensuring that the user experiences a gradual transition towards belief harmony. RGRec comprises three key modules, each integral to the elimination of "filter bubbles". Firstly, the **Multi-faceted Reasoning-based "filter bubbles" Detection module (FBDetect)** plays a pivotal role in our system by comprehensively detecting "filter bubbles". Consider the example in Fig. 4.1, where user "U21538" is identified by FBDetect as a filter bubble-affected user with extremely imbalanced beliefs. Specifically, FBDetect recognizes that this user highly favors "autos" while displaying minimal interest in "news". These precise insights about users with extremely imbalanced beliefs enable RGRec to implement targeted interventions to counteract the effects of "filter bubbles" effectively. Subsequently, the FBDetect module transfers these findings to the Belief Nudging module, which further contributes to achieving user belief harmony. The **Belief Nudging module** forms the core of RGRec. Its main task is to use the nudging strategy and topic similarities to provide a recommendation prompt path for the next module, RecomGen. As depicted in Figure 4.1, the module takes in two key inputs from the FBDetect: the user's preferred "autos" and the less preferred "news". The nudging strategy combines topic similarities to further generate the prompt path based on these two topics,

which is "autos → lifestyle → sports → news". This path acts as an input for the RecomGen. Finally, the **Recommendation Strategy Generation module (RecomGen)** is the "creative" component of RGRec, tasked with generating an array of items based on the prompt path, denoted as *GI*. Drawing on the path charted by the Belief Nudging module, it constructs recommendations that are relevant and varied, enriching the user's experience. The *GI* blends with the initial recommendation set, ensuring that the final recommendations delivered to the user are comprehensive and engaging. As users engage with these recommendations, their responses are fed back into the system to refine future recommendations.

Through these three core modules, RGRec systematically refines user beliefs and optimal paths in response to user interactions. Throughout this iterative process, the system continuously assesses the likelihood of user acceptance, adapting its approach until it aligns with the user's beliefs, thereby achieving harmony. Simultaneously, as RGRec expands the recommended items using the output of a preference-based recommendation system, it effectively upholds both responsibility and usability.

4.4.1 The Multi-faceted Reasoning-based "filter bubbles" Detection module (FBDetect)

The Multi-faceted Reasoning-based "filter bubbles" Detection module (FBDetect) is a module designed to identify "filter bubbles" and users with extreme beliefs from two perspectives: system-level bias and user belief bias. Unlike conventional single-dimensional reasoning models, FBDetect operates in two complementary modes: Forward Reconnaissance (FR) and Counter Reconnaissance (CR). These components work together to identify and confirm the presence of "filter bubbles". The FR component evaluates the recommendation system, assessing its potential for causing "filter bubbles". It explores system-level bias and behaviors to determine if the system perpetuates a filter bubble effect. In contrast, the CR component focuses on users, pinpointing those influenced by "filter bubbles" with extremely imbalanced beliefs. FBDetect offers a comprehensive evaluation of "filter bubbles", providing insights into biased recommendation models and users with extremely imbalanced beliefs. This implementation contributes to developing fairer recommendations and interventions to mitigate the adverse effects of "filter bubbles". For a visual representation of FBDetect's structure (see Fig. 4.2).

Forward Reconnaissance (FR) Component

For the FR component, I quantify the potential filter bubble influence of the current recommendation model using mathematical methods. The aspect coverage score [218] is employed

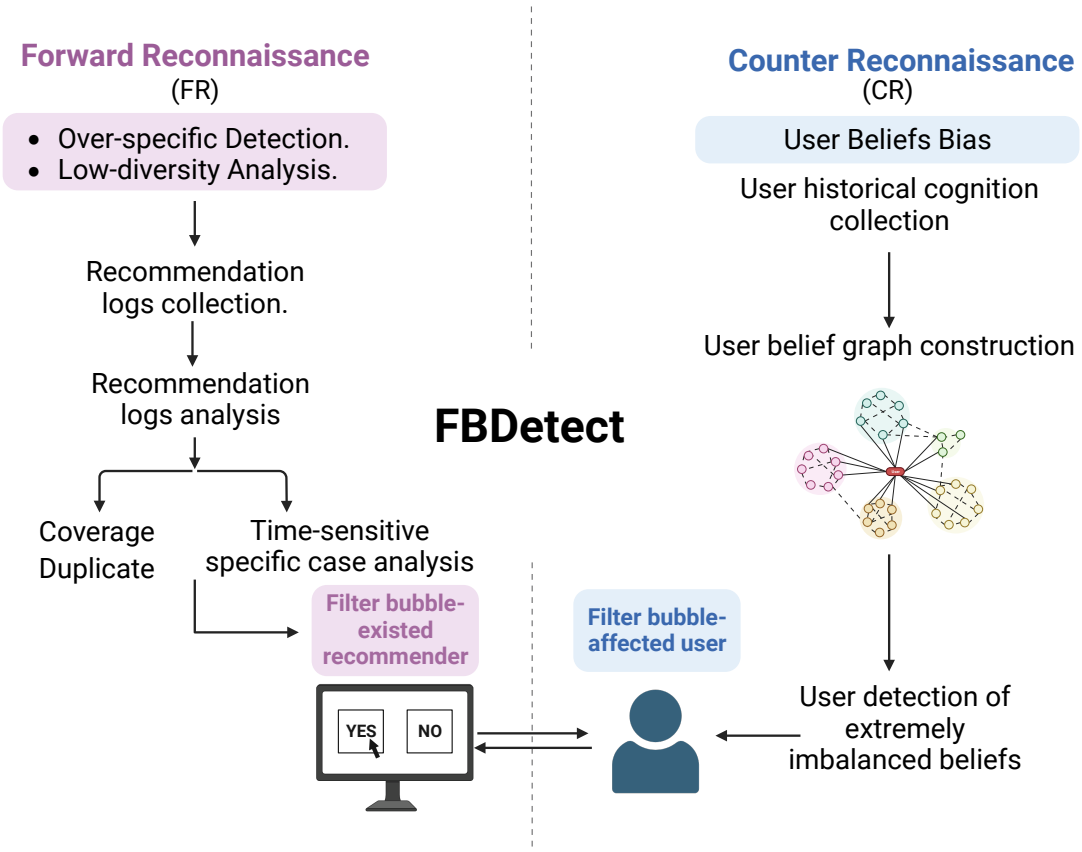


Fig. 4.2: FBDetect: The Multi-faceted Reasoning-based "filter bubbles" Detection module

in this FR component as a key mathematical validation metric. The formula for calculating aspect diversity coverage μ is given below:

$$\mu = \frac{|\{c_x | I_k^x \cap feed \neq \emptyset, \forall x, \forall k\}|}{|C|} \times \frac{|\{c_k^x | I_k^x \cap feed \neq \emptyset, \forall x, \forall k\}|}{\sum_{c_x \in C_{feed}} |C_x^{sub}|}. \tag{4.4}$$

The first part measures the diversity of topics by calculating the proportion of different topics present in the feed relative to the total number of topics available. While the second part measures the diversity of the aspects within those topics, where the proportion of different aspects covered in the feed for the topics that are present in the feed, relative to the total number of aspects in those topics.

Counter Reconnaissance (CR) Component

The CR component includes two key steps: constructing the user belief network and detecting users with extremely imbalanced beliefs.

Constructing the user belief network: This step involves creating a specific belief network for each user, derived from their historical interaction records. This belief network is employed to analyze user preferences toward different topics and identify users with extremely imbalanced beliefs. Fig. 4.3 illustrates a representative user belief network for the user "U21538", showcasing an example of user belief network construction within the CR component.

In Fig. 4.3, the red, blue, and yellow nodes represent the topics "news", "autos", and "lifestyle", with which user "U21538" has historically interacted. The pink nodes represent the aspects related to each topic that the user has interacted with. The weights over edges labeled "Includes", linking a topic to its aspects, are calculated based on the user's click probabilities towards these aspects. Additionally, the connection between the user and a topic, "belief", indicates the user's preference towards the topic and is calculated using the entropy metric in Equation 4.2.

The CR component constructs a specific user belief network for each user in our dataset, enabling clear identification of users' preferences. Its primary aim is to identify users with extremely imbalanced beliefs. I combine these constructed belief networks with the empirical rule (the "68-95-99.7" rule) [225] to explore users with extremely imbalanced beliefs.

Detecting users with extremely imbalanced beliefs: As previously mentioned, Fig. 4.3 shows the user belief network of "U21538". From this figure, I can deduce the user's preferences for "autos" and "news", which are 0.799 and 0.445, respectively. To determine whether this user has an extremely imbalanced belief, I collect statistics on all users' preferences for the "autos" and "news". As illustrated in Fig. 4.4, these statistics include users' preferences

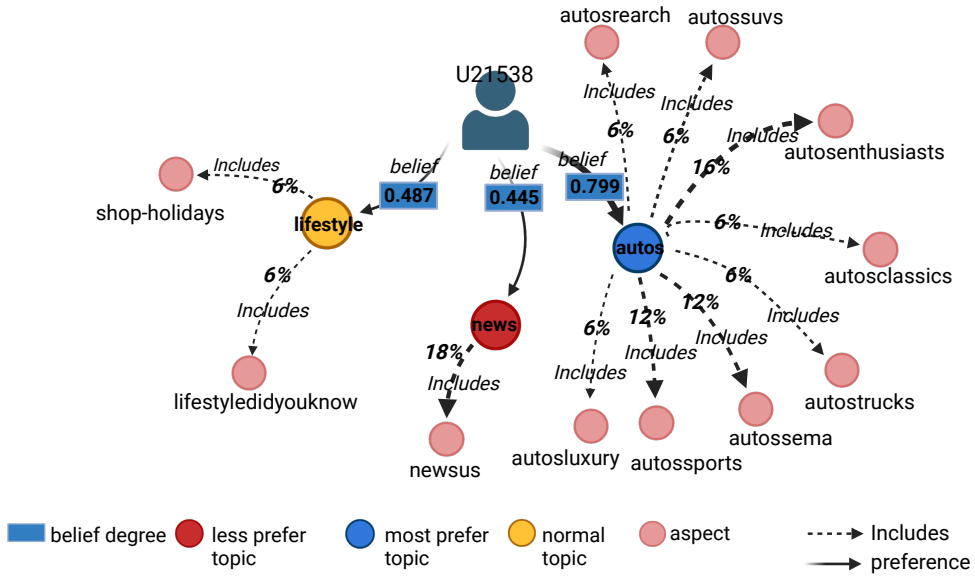


Fig. 4.3: The user belief network for user “U21538”.

for “autos”, with the x-axis representing the preference value and the y-axis representing the number of users. I identify “U21538” as a user with extremely imbalanced beliefs by following these steps:

- *Verify Normal Distribution:* The initial step involves verifying whether the preference distribution chart conforms to a normal distribution. To achieve this, I utilize the Kolmogorov-Smirnov (K-S) test [190]. In our scenario, the K-S test is a statistical method employed to compare the preference distribution of a specific topic with the normal distribution. The results of the K-S test typically include the value of the K-S statistic and its associated p-value. The p-value denotes the probability of observing the K-S statistic under the assumption that the distributions of the two datasets are identical. A small p-value (in this study, I set the threshold at 0.05) allows us to reject the null hypothesis that the two datasets share the same distribution, indicating statistically significant differences. Therefore, in Fig. 4.4, the p-value exceeds 0.05, suggesting that the preference distribution within “autos” conforms to a normal distribution.

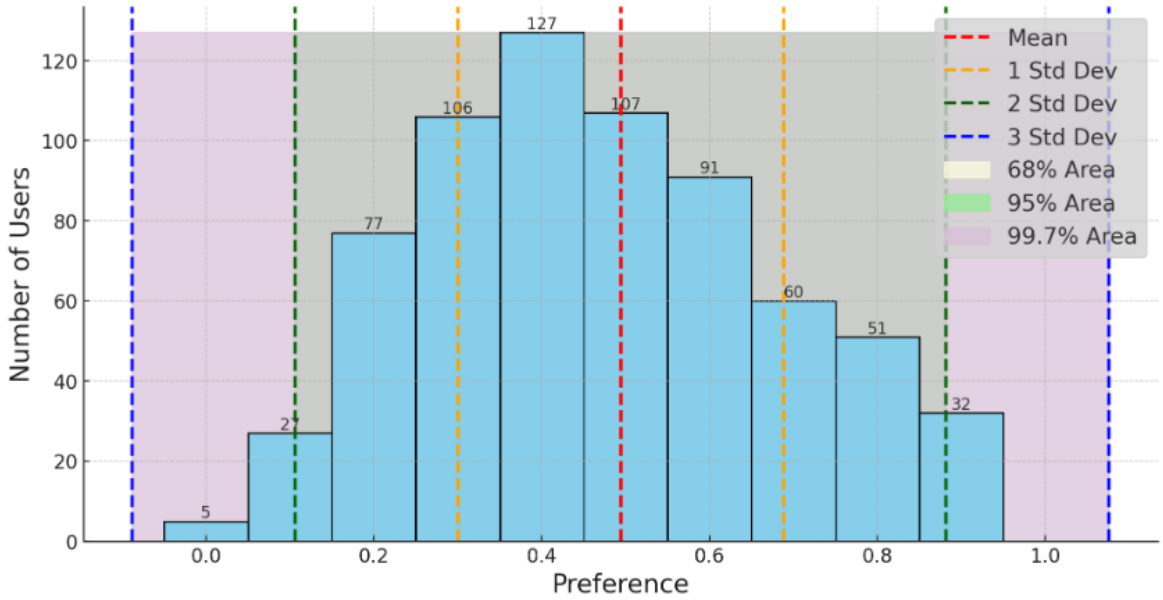


Fig. 4.4: A specific example of a preference distribution chart within "autos"

- *Distribution Segment:* Once it is confirmed that the preference distribution chart adheres to a normal distribution, the next step involves segmenting the distribution using the "68-95-99.7" rule [225], as depicted in Fig. 4.4. In the figure, it can be seen that approximately 68% of the data falls within one standard deviation of the mean, about 95% within two standard deviations, and roughly 99.7% within three standard deviations. In our study, I categorize users outside the 68% area (outside the yellow lines) as having extreme beliefs in the topic. Therefore, when user *U21538* (see Fig. 4.3) prefers "autos" with 0.799 degrees, falling outside the 68% area in Fig. 4.4, and shows less interest in "news" with 0.445 degrees, also falling outsides the 68% area in the "news" preference distribution chart. This user is identified as having extremely imbalanced beliefs and is further forwarded to the Belief Nudging module to mitigate the extremely imbalanced beliefs, i.e., "filter bubbles".

4.4.2 Belief Nudging Module

RGRec combines users' belief graphs and nudging techniques to gently stimulate users' interests in topics that were initially less preferred. The primary goal of the Belief Nudging module within the RGRec framework is to identify the most effective recommendation prompt path for the subsequent RecomGen. This is achieved by bridging the gap between the user's most favored and less preferred topics.

Adaptive Path Exploration Algorithms

The recommendation prompt path begins with topics that users highly favor (e.g., "autos" as shown in Fig. 4.3) and ends with those topics they are less inclined towards (such as "news" in Fig. 4.3). To connect these points, an adaptive path exploration algorithm discovers additional topics, forming a comprehensive recommendation prompt path and laying the groundwork for future nudge-based recommendations.

The nudge recommendation strategy in RGRec is based on this recommendation prompt path, where the adaptive path exploration algorithm automatically constructs the path based on user feedback. This forms a closed-loop feedback system that interlinks user feedback and system recommendations, dynamically adapting to deliver contextually appropriate prompts tailored to the user's current preferences.

The adaptive path exploration algorithm, inspired by the shortest path exploration algorithm known as *Dijkstra's algorithm* [17], starts from a central point, traverses neighboring points, and identifies the point with the highest weight as the starting point for the next step. I have proposed an enhanced version of this algorithm to accommodate the dynamic nature of user and topic relationships. This algorithm integrates the evolving user perceptions and topic relationships into the path discovery process. The objective is to discover a recommendation prompt path $p_{i,t}$ for an identified filter bubble-affected user u_i at time step t. This path incorporates contextually rich items into the recommendations, thus providing more diverse content and promoting belief harmony.

In the recommendation prompt path $p_{i,t}$ for the user u_i, $c_x^{SP_i}$ and $c_y^{LP_i}$ represent the user's most and less preferred topics, respectively, serving as the path's start and end points. The initial node of the path is denoted as $p_{i,t}(k) = c_x^{SP_i}$ with k=1. The selection of the subsequent node, the $(k+1)^{th}$ topic c_x from the set C, is determined by maximizing the following expression:

$$p_{i,t}(k+1) = \arg\max_{c_{x'}} \rho(c_x^{SP_i}, c_{x'}) + b_{ix'} * rej_{w.t}, \tag{4.5}$$

where $\rho(c_x^{SP_i}, c_{x'})$ calculates the topic similarity between $c_x^{SP_i}$ and $c_{x'}$. The term $b_{ix'}$ is the belief degree of user u_i towards topic $c_{x'}$ at the current time step, $rej_{w.t}$ is a rejection weight, typically set to 1.

I incorporate a tolerance threshold, denoted as θ, to modulate the parameter $rej_{w.t}$, emphasizing the significance of user feedback. Users impacted by "filter bubbles" tend to have their decisions heavily influenced by these bubbles. An item congruent with a user's beliefs is more likely to be accepted, whereas items not aligned with preferred topics often face acceptance challenges. To evaluate the effectiveness of a topic within a recommendation

prompt path, I monitor the frequency of rejections for topic-related items. If a user consistently rejects items from *GI* that are generated based on the recommendation prompt path, these items are deemed ineffective, leading to an assignment of a $rej_{w.t}$ value of -1. This approach adapts effectively to evolving user preferences. The exploration process concludes once the path $p_{i,t}$ includes the user's less preferred topic $c_y^{LP_i}$.

Utilizing the start and end nodes identified as "autos" and "news" in Fig. 4.5, and employing the adaptive path exploration algorithm, the recommendation prompt path at time t is formulated as $p_{i,t}$ = "autos $\rightarrow$ lifestyle $\rightarrow$ sports $\rightarrow$ news".

Nudging Strategy

Our nudging process incorporates incremental computing techniques [159] to enhance the efficiency of recommendation calculations. By breaking down the recommendation path into smaller segments, or sub-paths, the system recalibrates only the affected sub-path in response to a user's specific preference, instead of recalculating the entire path. This segmented approach is particularly effective for managing lengthy paths and surpasses the capabilities of traditional sequential recommendation systems. It not only reduces the number of recommendations needed but also increases overall efficiency. Once the path "autos $\rightarrow$ lifestyle $\rightarrow$ sports $\rightarrow$ news" is established, the nudging strategy is employed to refine and finalize the recommendation prompt path. For a clear illustration, both Figure 4.5 and Algorithm 1 demonstrate the recommendation process within the RKGRec nudging framework. This process includes generating recommendation paths through nudging, creating *GI* by the RecomGen, and subsequently reconstructing the user belief graph.

In Fig. 4.5, the identified optimal path is "autos $\rightarrow$ lifestyle $\rightarrow$ sports $\rightarrow$ news". This sequence forms the basis for the nudging strategy, which is fine-tuned based on user feedback. Conforming to the principles of incremental computing, the primary focus is initially on the "autos $\rightarrow$ lifestyle" segment. RGRec dynamically updates the user's belief network as the user positively engages with the "GI" content generated from RecomGen based on this segment. If the user's most and least preferred topics remain constant, the model progresses to recommend the subsequent segment, "sports $\rightarrow$ news", instead of restarting the entire path exploration process. However, if the user's topic preferences shift, these altered topics are reintroduced to the adaptive path exploration algorithm to create a new recommendation path.

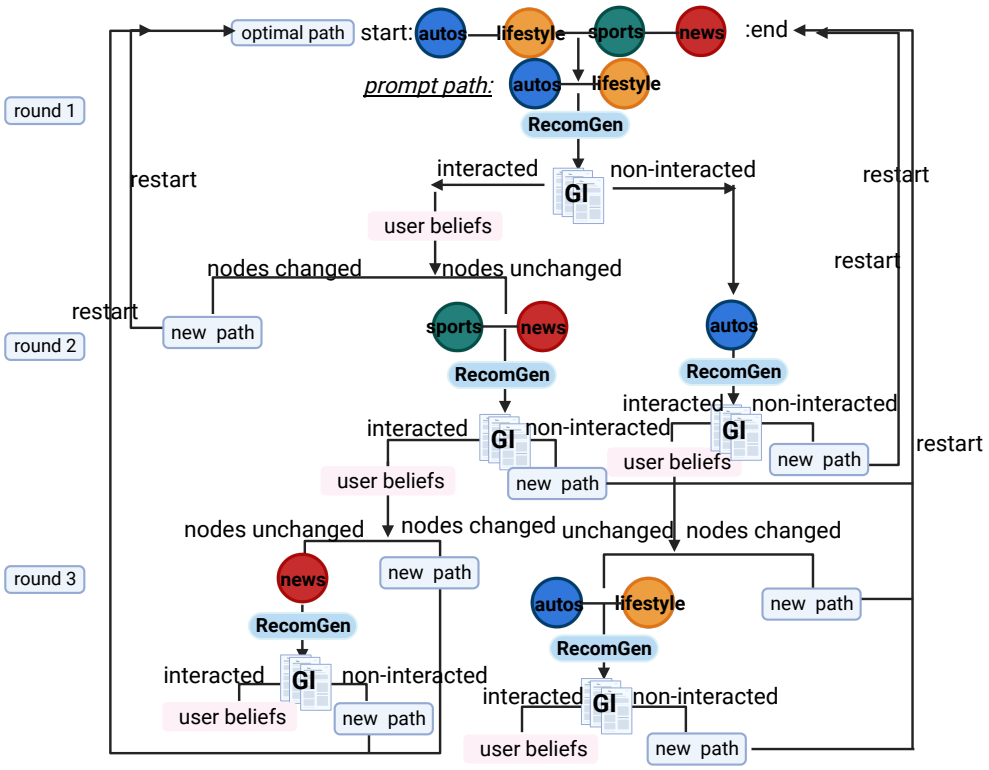


Fig. 4.5: An actual example of nudging recommendations process

4.4.3 The Generative Artificial Intelligence-based Recommendation Strategy Generation module (RecomGen)

When combined with nudge recommendation techniques, the RecomGen efficiently exploits the interconnectedness of information. This combination presents a solution to address information gaps that may arise during end-to-end recommendation processes. By employing the capabilities of the Large Language Models (LLM), e.g., GPT-3.5 Turbo, this approach offers rich semantic information at each step in the recommendation path, thereby strengthening the relationships between individual points. This strategy effectively engages users' interest in specific topics, fostering intrigue in areas they might find less attractive.

As previously mentioned, a nudging prompt $p_{i,t} = \{c_x^{SP_i}, ..., c_y^{LP_i}\}$ is generated for each time step. This prompt represents an optimal path between the starting node $c_x^{SP_i}$, and the end node $c_y^{LP_i}$. To leverage rich contextual information from point-to-point paths within the vast landscape of big data, these paths are inputted into the RecomGen as keywords or prompts. This generates contextually rich items *GI* based on the prompt path.

For example, in Fig. 4.5, the responsibility of the RecomGen is to obtain the path from each $feed$ and generate *GI* based on the path. In conclusion, integrating the RecomGen with the nudge strategy in RGRec effectively utilizes interconnections. This approach bridges information gaps, leverages the rich semantic information provided by RecomGen, and employs a nudge strategy to establish strong connections between data points. This strategy effectively stimulates user interest in topics that initially receive less attention and promotes user belief harmony.

Acceptance Probability Equation

The RecomGen incorporates a list of contextual rich items *GI* into the original recommendation list $feed_{original}$ and generates the final recommendation list $feed = \{i_1, i_2, .., i_j\}$. Once user u_i receives this $feed$, the probability $AP_{u_i}^{i_j}$ that whether to accept an item i_j in the recommendation list can be calculated using Equation 4.6:

$$AP_{u_i}^{i_j} = \frac{b_{ix}}{\sum_{c_{x'} \in \hat{C}_i} b_{ix'}}, i_j \in I_k^x \tag{4.6}$$

where b_{ix} represents the belief degree of u_i towards topic c_x, and $\sum_{c_{x'} \in \hat{C}_i} b_{ix'}$ calculates the total belief degrees of u_i.

4.5 Experiments

I conduct experiments from two general directions: system and user. The system-centered experiment primarily aims to demonstrate the effectiveness of RGRec as an intermediate agency in alleviating the system filter bubble. From the user's perspective, four user-centered experiments are conducted. These include detecting RGRec's positive effect on increasing user belief diversity, examining its effectiveness in motivating filter bubble-impacted users' interests in topics they are initially less interested in, and analyzing RGRec's ability to reduce the number of filter bubble-impacted users. Finally, I perform two parametric analyses to assess the effects of different RGRec recommendation weights and explore the impact of the threshold in RGRec on user belief diversity.

4.5.1 Experiment Setup

Dataset

In the experiments, I utilize two real-world datasets: the Microsoft News Dataset (MIND) [1] and IMDB [2] Dataset. MIND is a public news recommendation dataset encompassing user interaction data gathered from Microsoft News. It comprises data from 5,000 users, encompassing 230,117 user reading records and 51,287 news with 17 topics. IMDB is a movie recommendation dataset consisting of 25,000 movie rating records from 333 users and a selection of 2,586 movies in 16 movie topics. I adopt a pre-trained language model, BERT [108], to represent textual features in a vector space, capturing the semantic essence of the items. Such a method has already learned much about language structures and patterns [89] and has been widely adopted in recommendation studies [250, 98, 157].

The FBDetect module in RGRec, identified 180 and 20 filter bubble-affected users from the MIND and IMDB datasets, respectively. These identified users, denoted as u^{FB}, are the focus of subsequent filter bubble mitigation experiments.

Simulation of User Behaviors.

Given the impracticality and high cost associated with online testing for researchers, I have designed an offline evaluation approach: (1) Implement an "Acceptance Probability Equation" 4.6 to simulate user feedback, (2) generate recommendations using a "Nudge Strategy" based on the simulated user feedback, and (3) evaluate the recommendations by considering diversity and efficacy.

[1] https://msnews.github.io/

[2] https://www.kaggle.com/datasets/meastanmay/imdb-dataset?select=tmdb_5000_movies.csv/

4.5.2 Parameter settings and baselines

Baselines: I assess the performance of RGRec in comparison with several established baseline methods:

- Content-Based Filtering (CBF) [162]: This strategy recommends existing items based solely on content-based filtering.
- Collaborative Filtering (CF) [204]: This approach recommends existing items using only user collaborative filtering.
- Neural Graph Collaborative Filtering (NGCF) [221]: Enhances recommendation by using user-item graphs to model collaborative signals.
- LightGCN (LGCN) [91]: A simplified, yet effective, model focusing on neighborhood aggregation; outperforms NGCF while being easier to train.
- Disentangled Graph Collaborative Filtering (DGCF) [222]: Captures user intent diversity by analyzing user-item relationships.
- RGRec: This method suggests a set of items (designated as *GI*) as recommendation feeds, derived using the RGRec approach.

Building upon the baselines described above, in our experiments, I combined RGRec with each of the baselines (except the standalone RGRec), producing six additional experimental baselines. These combinations are all named with the superscript "$*$", such as CBF* and LGCN*. In total, I used 11 baselines for comparative experiments, aiming to demonstrate the superior performance of the models when combined with RGRec in breaking the filter bubble and moderating user extreme beliefs.

Evaluation Metrics:

I assess RGRec from both system and user perspectives. From the system viewpoint, Experiment 4.5.3 measures recommendation diversity using aspect coverage μ as defined in Equation 4.4. To confirm that observed differences in experimental outcomes are not due to chance, I employ the Two-Tailed Test t and corresponding p-value [175], attributing significant differences to model factors. The value of t is shown below:

$$t = \frac{\bar{X_1} - \bar{X_2}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{4.7}$$

where $\bar{X_1}$ and $\bar{X_2}$ are the means of the two sample groups, and S_p is the pooled standard deviation. n_1 and n_2 represent the sample sizes of these groups. The p value represents the

probability of observing the t statistic observed, assuming no significant difference exists between the two sample groups.

From the user perspective, our experiments focus on three areas. In Experiment 4.5.3, I evaluate the diversity of users' belief networks using the same μ metric in Equation 4.4. Experiment 4.5.3 analyzes the evolution of users' interests in highly favored and less favored topics over time, employing Equation 4.2. Lastly, Experiment 4.5.3 involves counting the number of filter bubble-affected users, using the CR of the FBDetect module, as detailed in Section 4.4.1.

Parameters: The proportion w of RGRec-generated items *GI* is considered as a parameter in our experiments. I analyze the impact of varying proportions w within a recommendation $feed$ on user belief and the diversity of the recommendation system. Additionally, I conduct a parameter analysis experiment of the tolerance threshold θ, which involves tracking user feedback regarding the generated *GI*, described in Section 4.4.2.

4.5.3 Experimental Results

Experiment 1: Coverage Analysis

The primary aim of this experiment is to assess the impact of RGRec on content diversification compared to the baseline model, and its effect on user belief networks. To accomplish this, I identified users affected by "filter bubbles" using specific criteria, including users such as "U25354" from the MIND dataset and "U128" from the IMDB dataset. For each of these users, I conducted a series of experimental rounds where they received 10 recommendations in each round, corresponding to a unique "feed". This process was repeated for 100 rounds. The repetition of these rounds aimed to ensure the reliability of our results by reducing the influence of randomness. Finally, the data from these 100 rounds were averaged to obtain the final experimental results, offering a comprehensive evaluation of RGRec's effectiveness in enhancing diversity in recommendations.

Coverage Analysis for Systems: Initially, I analyzed the evolution of content diversity across seven different models, focusing on selected users (user "U25354" from MIND and user "U128" from IMDB). The detailed results are presented in Table 4.2, showcasing each model's diversity coverage. Note that "*sum.*" represents each model's total diversity coverage degree throughout the recommendation process, and "*Improv.*" indicates the growth rate in diversity.

In the MIND dataset, the models with *wR* suffix, indicating RGRec integration, generally exhibit significant improvements in diversity coverage compared to their standard counterparts. Notably, models such as DGCF*, NGCF*, and LGCN* show substantial enhancements,

as reflected by their high *sum.* values and growth rates. This indicates that incorporating RGRec markedly enhances recommendation diversity.

Similarly, in the IMDB dataset, models integrated with RGRec demonstrate considerable improvements in diversity. The impact of RGRec is particularly pronounced in models like NGCF* and LGCN*, which display the highest growth rates in diversity.

The statistical significance values for all models integrated with RGRec on both datasets are all below the 0.05 threshold compared with the base model, suggesting that the improvements in diversity coverage by RGRec-integrated models are statistically significant compared to their counterparts. This underscores the vital role of RGRec in augmenting recommendation diversity.

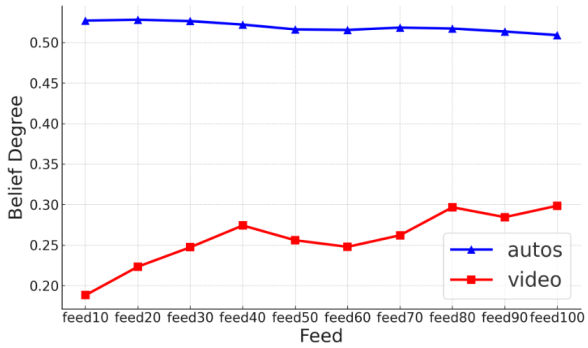
Overall, the analysis demonstrates that integrating RGRec into recommendation models significantly enhances the diversity of content recommended to users, which is crucial for mitigating "filter bubble" effects.

Coverage Analysis for User Belief Networks. Table 4.3 offers a detailed examination of the impact of different recommendation models on the diversity of user beliefs across the MIND and IMDB datasets.

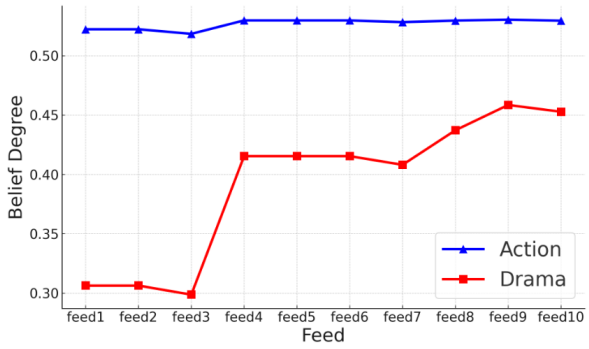
In both datasets, models integrated with RGRec demonstrate a significant improvement in total diversity coverage (*sum.*) compared to their respective baseline models. This indicates that integrating RGRec effectively broadens the range of user beliefs. The most notable diversity gains are seen in models like DGCF*, NGCF*, and LGCN*, especially in the MIND dataset. This highlights the effectiveness of these RGRec-enhanced models in diversifying user recommendations.

The *Improv.* metric shows a marked percentage increase in user belief diversity for the RGRec integrated models, particularly in the MIND dataset. For instance, the DGCF* model shows improvements exceeding 600%. The consistent enhancements across various models and both datasets reinforce RGRec's effectiveness in enhancing user belief diversity. The statistical significance values for all models integrated with RGRec on both datasets are all below the 0.05 threshold compared with the base model, confirming that the improvements in user belief diversity are statistically significant and attributable to the model variations, particularly due to RGRec integration. This statistical robustness emphasizes the reliability of the trends observed.

Overall, the analysis confirms that the inclusion of RGRec markedly enhances the diversity of user beliefs, which is crucial in mitigating "filter bubble" effects and enriching user belief diversity.



(a) Temporal Variation in User Beliefs about Topics of Most Interest and Less Interest on MIND



(b) Temporal Variation in User Beliefs about Topics of Most Interest and Less Interest on IMDB

Fig. 4.6: Temporal Variation in User Beliefs about Topics of Most Interest and Less Interest

Experiment 2: User Beliefs Analysis

RGRec is designed to present diverse information to users based on existing preference-based recommendation systems. Our goal is to demonstrate that users may come to accept information they initially preferred less through RGRec recommendations over a relatively short timeframe, thereby broadening their perspectives. To this end, I selected user "U276" from the IMDB dataset, who presents a shorter recommendation path $p_{shorter.u}$, indicating closer feature distances between most and less interested topics. This approach allows us to observe changes in user beliefs within a short period. Additionally, user "U18469", with a longer recommendation path $p_{longer.u}$, is included in the experiment to represent cases with longer recommendation paths.

This experiment employs the CF* model to investigate whether users, in the context of RGRec, will shift their focus from their highly preferred topic to a topic they are less interested in. Fig. 4.6 illustrates the temporal evolution of interest levels in topics initially less interesting and highly interesting to the users, as indicated by their belief networks.

For the user with longer paths, "p_{longer}", I have adjusted the timescale in Fig. 4.6 so that a single time interval now represents 10 recommendation feeds. This allows us to measure the user's interest level in the "autos" and "video" every 10 recommendations. In contrast, for the user with shorter paths, "$p_{shorter}$", I track belief changes after each recommendation. In these figures, "autos" and "Action" refer to the topics of high interest, while "travel" and "Drama" represent the less favored topics for each user.

From both figures, it is observed that the users' preferences for "autos" and "Action" remain relatively stable or exhibit a slight decline over time. On the other hand, continuous recommendations progressively increase their interest in the initially less favored topics, such as "video" and "Drama". This trend is particularly pronounced for the IMDB user, who shows a significant increase in interest in "Drama". These observations robustly validate the effectiveness of the RGRec-based model in promoting belief harmony among users by diversifying their interests.

Experiment 3: Filter Bubble Users Detection

Table 4.4 provides a comprehensive overview of the filter bubble effect across different recommendation models in both the MIND and IMDB datasets. This experiment focuses on tracking the changes over time in the number of users influenced by the filter bubble effect.

In the MIND dataset, a consistent pattern is observed where models integrated with RGRec consistently affect fewer users with the filter bubble than their original counterparts. This trend is evident across all recommendation feeds, showcasing RGRec's effectiveness in reducing the filter bubble impact.

Particularly, the original DGCF and NGCF models, which initially show a high number of users affected, see a significant decrease following RGRec integration. This highlights the model's capability to diversify user recommendations.

Similarly, in the IMDB dataset, RGRec-enhanced models also demonstrate a reduction in the number of users affected by the filter bubble. For example, the CF* model shows a decrease from 6 to 4 users affected and then exhibits fluctuation, reflecting the dynamic nature of user preferences and the system's adaptability. RGRec's ability to adjust the recommendation path based on user feedback likely contributes to these changes, allowing for a more responsive recommendation system that aligns with evolving user interests.

Overall, the analysis from the table confirms the effectiveness of the RGRec strategy in mitigating the filter bubble effect across different recommendation models and datasets. The observed temporal fluctuations further highlight the dynamic interaction between users and the recommendation system.

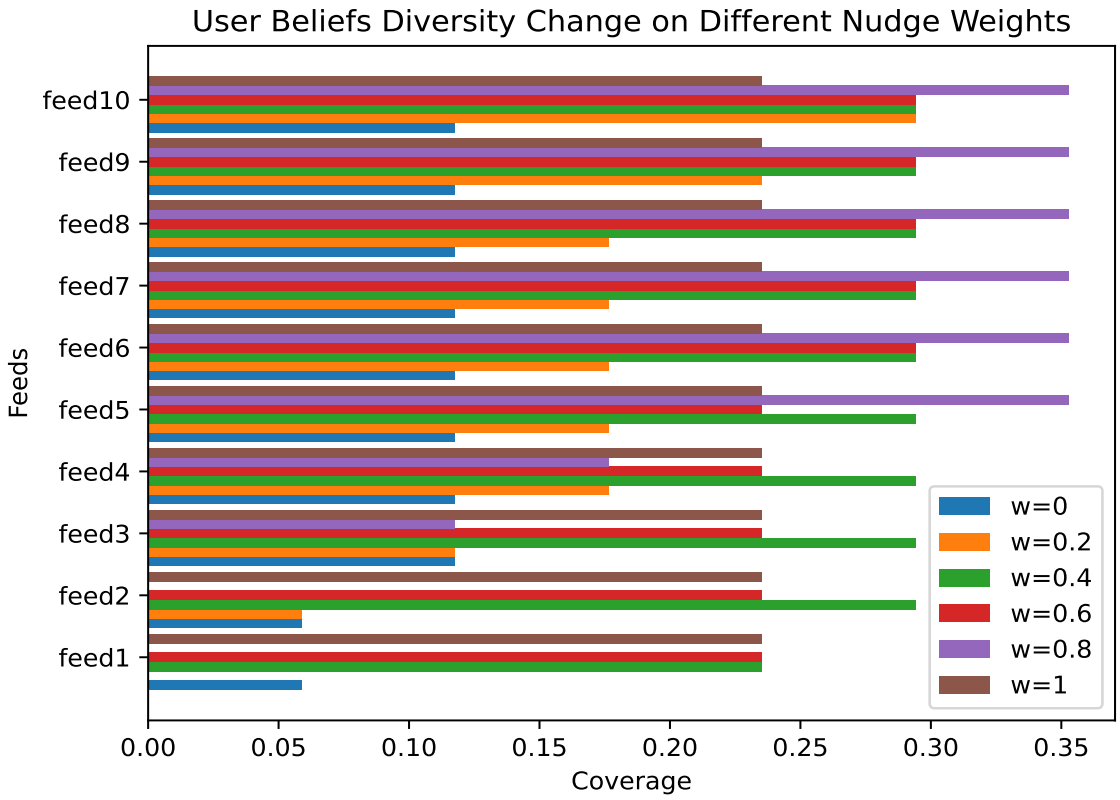


Fig. 4.7: User Beliefs Diversity Change on Different Nudge Weights

Experiment 4: Parameter Analysis-1

As mentioned in Subsection 4.5.2, RGRec utilizes two key parameters: the proportion w of RGRec-generated contextually rich items GI, and a tolerance threshold θ that tracks user feedback on these items. This experiment examines the impact of varying the weights w of RGRec-generated items within a recommendation feed on user belief diversity. I focus on how changing the weight of RGRec-generated items influences the diversity of user perspectives. By adjusting the w parameter, I assess the extent of RGRec's influence on the recommendation outcomes. The experiment employs user "U1629" from the MIND dataset as a case study and uses the CF* model.

As shown in Fig. 4.7, the weight assigned to RGRec recommendations has a significant effect on user belief diversity. It becomes clear that as the number of recommendations increases over time, the impact of the recommendation weight grows more pronounced. Notably, when the weight is set at 0.4, users are more likely to accept the recommendations, and this influence remains steady. However, after a certain number of recommendations, the impact of RGRec-recommended information on users reaches its peak. The more significant

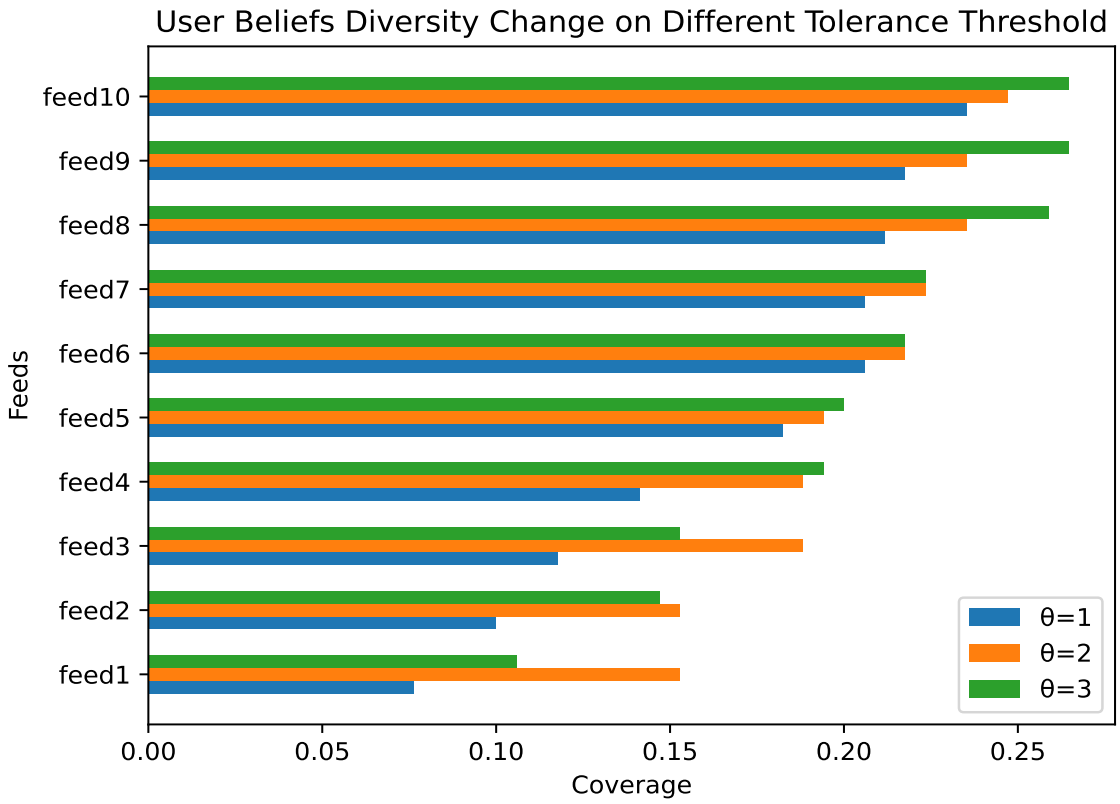


Fig. 4.8: User Beliefs Diversity Change on Different Tolerance Threshold

the fraction of items recommended by RGRec in the total recommendation list, the more significant the impact on users in the later stages of recommendations.

In conclusion, the analysis shows that the weight of RGRec recommendations considerably affects user belief diversity. The influence of this weight amplifies as the number of recommendations increases. However, there is a saturation point at which the impact of RGRec-recommended information on users reaches its maximum. In the experiment, I select w=0.6 as the RGRec recommendation weight owing to its consistently improving performance.

Experiment 4: Parameter Analysis-2

In this experiment, I investigate the impact of the tolerance threshold in RGRec on user belief diversity. The threshold setting reflects the model's consideration of user feedback, determining how many times a user must reject recommendations before RGRec alters its nudge recommendation strategy. I simulate various scenarios where users reject recommendations to mimic different feedback situations.

Analyzing Fig. 4.8, it is clear that user belief diversity gradually increases as the number of recommendations grows, with different threshold values playing a pivotal role. Higher thresholds result in greater user belief diversity, showcasing the system's enhanced adaptability to user feedback. Notably, when the threshold value is 3, the user belief diversity achieves the largest value, surpassing 0.25. This implies that a moderate threshold encourages the system to adjust its recommendations based on user feedback while still maintaining substantial diversity. The experiment intentionally avoids setting thresholds higher than 3 to prevent persistently recommending items that users dislike, which could lead to disengagement from the recommendation system. Across all thresholds, initial acceptance of recommendations is typically low. However, user acceptance of RGRec gradually increases as it adapts to feedback from multiple rejections, refining its nudge strategy and enhancing the overall effectiveness of the approach.

In conclusion, this experiment highlights the critical role of threshold settings in influencing user belief diversity. The findings suggest that a threshold value of 3 maximizes user belief diversity while appropriately respecting user preferences.

4.6 Discussion

The experiments illustrate that models integrating RGRec as an intermediary significantly outperform those without RGRec. RGRec is particularly potent in reducing "filter bubbles" and achieving a more balanced user belief network. It proves to be more effective and efficient in diversifying recommendations and fostering increased user belief diversity as the recommendation count grows.

Furthermore, the experiments assessed the impact of RGRec on users with varying weights of RGRec recommendations. The findings reveal its effectiveness in increasing users' acceptance of initially less preferred information, particularly for those with shorter recommendation paths. For users with longer paths, the model aids in gradually adapting to less favored information, underscoring its capacity to improve information acceptance and balance user beliefs.

In summary, experimental results demonstrate that RGRec excels in mitigating filter bubble effects, underscoring its substantial contribution to the evolution of recommendation systems. Beyond just enhancing recommendation diversity, RGRec fosters belief harmony, carrying meaningful implications for the elevation of user satisfaction. These insights play a pivotal role in guiding the development and refinement of future personalized recommendation models and strategies.

4.7 Conclusion and Future Work

In this chapter, I focused on addressing the issue of "filter bubbles" in RSs and proposed a Responsible Graph-based Recommendation system (RGRec) as a solution to mitigate the negative effects of "filter bubbles" by promoting belief harmony among users.

RGRec acts as an intermediary between existing preference-based recommendation systems and users, aiming to facilitate more democratic recommendations. The model features several core components: FBDetect, for filter bubble identification; nudging techniques to gradually broaden users' interests and balance their beliefs; and a user feedback loop based on RecomGen, which tracks evolving user beliefs over time and enhances the diversity of recommendations.

The experimental results clearly show the effectiveness of RGRec in mitigating "filter bubbles" and fostering a more balanced belief system among users. However, the challenge remains in continuously refining the system to better serve users' evolving needs.

As we transition to the next chapter, the focus shifts from the specific challenge of mitigating filter bubbles to exploring a broader framework for balancing personalized information delivery with neutrality. This introduces an agent-based approach designed to ensure the responsible flow of information, promoting not just diversity but also fairness and neutrality in recommendations. Moreover, the chapter will delve into the design of a user-centered system that effectively balances accuracy, diversity, and fairness, establishing a comprehensive evaluation framework. This will further broaden the responsible recommendation framework by ensuring that the system is adaptable to users' needs and societal expectations, ultimately fostering a well-rounded, equitable user experience.

Table 4.1: Table of Notations

Notation	Description
S	The recommendation environment.
U	A set of users.
A	AI-based algorithm for generating recommendations.
C	A set of predefined topics.
u_i	A user u_i in the system.
c_x	A topic c_x
C_x^{sub}	A set of aspects associated with a topic c_x.
c_k^x	An aspect associated with c_x.
I_k^x	A set of items whose aspect is c_k^x.
i_m	An item.
G_i	Belief network of user u_i.
V_i	A set of nodes in G_i.
$\hat{C}_i$	A set of topics that u_i interacted with.
$\hat{C}_i^{sub}$	A set of aspects that u_i interacted with.
E_i	A set of edges in G_i.
E_i^b	A set of edges connecting from u_i to topics.
E_i^c	A set of edges connecting from topics to aspects.
e_{ix}	An edge from u_i to c_x.
e_{xk}	An edge from c_x to c_k^x.
b_{ix}	Belief degree associated on edge e_{ix}.
r_{ik}^x	Click probability associated on edge e_{xk}.
$\rho(c_x, c_y)$	Topic similarity measure between two topics.
$p_{i,t}$	Recommendation prompt path for user u_i at time step t.
$p_{i,t}(k)$	The k^{th} node in the recommendation prompt path $p_{i,t}$.
$feed$	A recommendation list comprising $feed_{original}$ and GI.
C_{feed}	A subset of C that includes only those topics appeared in the feed.
$feed_{original}$	A list of items suggested by the preference-based recommendation system.
GI	Contextually rich items generated by RGRec, based on an explored recommendation prompt path.
$\check{M}_{i,t}$	A set of items within a $feed$ accepted by user u_i from the initial time step to time step t.
$\bar{P}_{i,t}$	A sequence of recommendation prompt paths shown to u_i, where resulting items have been declined over the same period.

Algorithm 1 Nudge Recommendation Process of RGRec

Input: the current recommendation prompt path $p_{i,t} = \{c_x^{SP_i}, ..., c_y^{LP_i}\}$ as the current prompt p

Initialise an empty heap $Q = []$
Q = Binary Split Function(p) ▷ Binary Split Function adopts incremental computing techniques to break down path p
while length(Q) > 0 **do**
    Set current prompt $p = Q[0]$
    Generate contextually rich items GI with prompt $p = Q[0]$
    Recommend GI to user u_i
    if u_i accepts GI **then**
        $\check{M}_{i,t} = \check{M}_{i,t} \cup$ GI ▷ $\check{M}_{i,t}$ represents user u_i's collection of accepted items until time step t
        Update user belief graph G_{u_i}
        Q.pop(0)
    else
        $\bar{P}_{i,t}.append(p)$ ▷ $\bar{P}_{i,t}$ records user u_i's declined recommendation prompt paths from time 0 to t
        if count$(\bar{P}_{i,t}, p) > \theta$ **then**
            Update rejection weight $rej_{w.t}$
        end if
        $p_t^1, p_t^2 =$ Binary Split Function (p)
        Q.pop(0)
        Push p_t^1, p_t^2 to Q
    end if
    if Binary Split Function (p) is None **then**
        Reschedule path
    end if
end while
function BINARY SPLIT FUNCTION(p_t)
    if $length(p_t) == 2$ **then**
        return
    else
        if $mod(length(p_t), 2) == 1$ **then**
            $p_t^1 = p_t[0 : \frac{length(p_t)-1}{2}]$
            $p_t^2 = p_t[\frac{length(p_t)+1}{2} :]$
        else
            $p_t^1 = p_t[0 : \frac{length(p_t)}{2}]$
            $p_t^2 = p_t[\frac{length(p_t)}{2} :]$
        end if
    end if
    return p_t^1, p_t^2
end function

Table 4.2: Coverage Analysis of Recommendation Models based on MIND and IMDB datasets. Boldface denotes the highest score. Marking with underline denotes the significance p-value$<$0.05 compared with the base model.

Times	MIND											IMDB										
	CB	**CB***	**CF**	**CF***	**DGCF**	**DGCF***	**NGCF**	**NGCF***	**LGCN**	**LGCN***	**RGRec**	**CB**	**CB***	**CF**	**CF***	**DGCF**	**DGCF***	**NGCF**	**NGCF***	**LGCN**	**LGCN***	**RGRec**
$feed_1$	0.058	0.294	0.184	0.335	0.176	0.294	0.294	0.470	0.294	0.412	0.294	0.062	0.500	0.212	0.526	0.188	0.438	0.250	0.375	0.125	0.479	0.500
$feed_2$	0.059	0.205	0.186	0.202	0.176	0.353	0.235	0.470	0.235	0.412	0.195	0.062	0.345	0.207	0.386	0.125	0.500	0.250	0.563	0.250	0.354	0.344
$feed_3$	0.058	0.154	0.181	0.207	0.176	0.176	0.294	0.294	0.294	0.294	0.132	0.062	0.238	0.209	0.328	0.125	0.469	0.250	0.375	0.125	0.250	0.216
$feed_4$	0.059	0.145	0.179	0.241	0.117	0.235	0.294	0.294	0.294	0.264	0.123	0.062	0.180	0.201	0.274	0.125	0.313	0.219	0.281	0.125	0.188	0.147
$feed_5$	0.059	0.151	0.182	0.215	0.117	0.117	0.294	0.294	0.294	0.294	0.119	0.062	0.163	0.207	0.259	0.281	0.344	0.219	0.281	0.188	0.250	0.127
$feed_6$	0.059	0.199	0.184	0.242	0.059	0.117	0.294	0.294	0.235	0.265	0.171	0.062	0.163	0.216	0.254	0.250	0.281	0.188	0.313	0.125	0.188	0.127
$feed_7$	0.059	0.205	0.176	0.212	0.117	0.117	0.235	0.412	0.294	0.294	0.174	0.062	0.190	0.209	0.268	0.188	0.250	0.219	0.344	0.125	0.250	0.149
$feed_8$	0.059	0.178	0.186	0.215	0.176	0.117	0.235	0.353	0.235	0.324	0.154	0.062	0.184	0.197	0.269	0.125	0.156	0.219	0.219	0.250	0.250	0.158
$feed_9$	0.059	0.164	0.185	0.224	0.117	0.117	0.294	0.471	0.294	0.382	0.139	0.062	0.170	0.214	0.247	0.062	0.188	0.188	0.250	0.125	0.188	0.139
$feed_{10}$	0.059	0.156	0.191	0.221	0.117	0.117	0.294	0.412	0.235	0.352	0.136	0.062	0.164	0.218	0.243	0.125	0.125	0.219	0.219	0.125	0.1875	0.135
sum.	0.588	**1.851**	1.832	**2.315**	1.352	**1.941**	2.765	**3.765**	2.706	**3.294**	1.501	0.625	**2.297**	2.089	**3.053**	1.594	**3.063**	2.219	**3.219**	1.563	**2.583**	2.041
Improv.	-	214.79%	-	26.31%	-	43.48%	-	36.17%	-	21.74%	-	-	267.6%	-	46.15%	-	92.16%	-	45.07%	-	65.33%	-

Table 4.3: Coverage analysis of user beliefs on the MIND and IMDB datasets. Boldface denotes the highest score. Marking with underline denotes the significance p-value$<$0.05 compared with the base model.

Times	MIND											IMDB										
	CB	**CB***	**CF**	**CF***	**DGCF**	**DGCF***	**NGCF**	**NGCF***	**LGCN**	**LGCN***	**RGRec**	**CB**	**CB***	**CF**	**CF***	**DGCF**	**DGCF***	**NGCF**	**NGCF***	**LGCN**	**LGCN***	**RGRec**
$feed_1$	0.056	0.090	0.056	0.104	0.059	0.294	0.059	0.294	0.059	0.294	0.047	0.038	0.097	0.049	0.090	0.062	0.375	0.031	0.375	0.062	0.062	0.085
$feed_2$	0.059	0.116	0.075	0.133	0.059	0.471	0.059	0.294	0.059	0.353	0.074	0.054	0.147	0.082	0.159	0.062	0.375	0.094	0.375	0.062	0.062	0.104
$feed_3$	0.059	0.126	0.089	0.161	0.059	0.471	0.059	0.294	0.059	0.353	0.091	0.059	0.172	0.111	0.198	0.062	0.375	0.094	0.375	0.062	0.062	0.126
$feed_4$	0.059	0.135	0.099	0.186	0.059	0.471	0.059	0.294	0.059	0.353	0.106	0.062	0.183	0.133	0.223	0.125	0.375	0.094	0.375	0.062	0.125	0.145
$feed_5$	0.059	0.144	0.110	0.204	0.059	0.471	0.059	0.294	0.059	0.353	0.116	0.062	0.191	0.151	0.243	0.125	0.375	0.125	0.375	0.062	0.125	0.159
$feed_6$	0.059	0.158	0.122	0.221	0.059	0.471	0.059	0.294	0.059	0.353	0.132	0.062	0.197	0.172	0.257	0.125	0.375	0.125	0.375	0.062	0.125	0.164
$feed_7$	0.059	0.165	0.131	0.228	0.059	0.471	0.059	0.294	0.059	0.353	0.142	0.062	0.208	0.189	0.279	0.125	0.375	0.125	0.375	0.062	0.125	0.174
$feed_8$	0.059	0.171	0.138	0.238	0.059	0.471	0.059	0.353	0.059	0.393	0.148	0.062	0.219	0.203	0.286	0.125	0.375	0.125	0.375	0.062	0.125	0.181
$feed_9$	0.059	0.178	0.149	0.245	0.059	0.471	0.059	0.353	0.059	0.393	0.155	0.062	0.228	0.213	0.300	0.125	0.375	0.125	0.375	0.062	0.125	0.188
$feed_{10}$	0.059	0.184	0.160	0.252	0.059	0.471	0.059	0.353	0.059	0.393	0.167	0.062	0.233	0.220	0.310	0.125	0.375	0.125	0.375	0.062	0.125	0.192
sum.	0.585	**1.467**	1.128	**1.970**	0.588	**4.529**	0.588	**3.112**	0.590	**3.588**	1.178	0.587	**1.873**	1.522	**2.344**	1.063	**3.750**	1.063	**3.750**	0.625	**1.062**	1.517
Improv.	-	150.75%	-	74.56%	-	669.99%	-	430.00%	-	510.00%	-	-	218.93%	-	53.98%	-	252.94%	-	252.94%	-	70.00%	-

Times	MIND											IMDB										
	CBF	**CBF***	**CF**	**CF***	**DGCF**	**DGCF***	**NGCF**	**NGCF***	**LGCN**	**LGCN***	**RGRec**	**CBF**	**CBF***	**CF**	**CF***	**DGCF**	**DGCF***	**NGCF**	**NGCF***	**LGCN**	**LGCN***	**RGRec**
$feed_1$	2	2	28	24 ↓	26	18 ↓	26	16↓	20	16 ↓	26	0	0	6	4 ↓	6	5 ↓	6	5 ↓	6	5 ↓	4
$feed_2$	2	2	28	24 ↓	26	16 ↓	26	16↓	20	16 ↓	26	0	0	6	4 ↓	6	5 ↓	6	5 ↓	6	5 ↓	4
$feed_3$	2	2	28	24 ↓	26	14 ↓	26	18↓	20	18 ↓	26	1	0 ↓	6	4 ↓	6	5 ↓	8	5 ↓	7	5 ↓	4
$feed_4$	4	4	28	24 ↓	26	14 ↓	28	18↓	22	18 ↓	26	2	0 ↓	6	6	8	5 ↓	8	5 ↓	7	5 ↓	4
$feed_5$	4	4	28	24 ↓	26	10 ↓	28	18↓	24	20 ↓	26	3	1 ↓	6	6	8	5 ↓	8	5 ↓	7	5 ↓	4
$feed_6$	4	4	28	24 ↓	26	10 ↓	28	18↓	26	18 ↓	26	3	2 ↓	6	6	8	5 ↓	8	5 ↓	7	5 ↓	4
$feed_7$	4	4	28	24 ↓	28	10 ↓	28	18↓	26	18 ↓	26	3	3	6	6	8	5 ↓	8	5 ↓	8	5 ↓	4
$feed_8$	6	4 ↓	28	24 ↓	28	10 ↓	28	16↓	28	18 ↓	24	3	2 ↓	6	6	8	5 ↓	8	5 ↓	8	5 ↓	4
$feed_9$	6	4 ↓	28	24 ↓	28	10 ↓	28	16↓	28	18 ↓	24	3	2 ↓	6	6	8	5 ↓	8	5 ↓	8	5 ↓	4
$feed_{10}$	10	4 ↓	28	24 ↓	28	10 ↓	28	16↓	28	16 ↓	24	4	2 ↓	6	5 ↓	8	5 ↓	8	5 ↓	8	5 ↓	4

Table 4.4: Filter Bubble Users Detection on MIND and IMDB datasets

Chapter 5

Responsible Balance in Information Delivery: An Agent-Based Neutrality Model for Recommendations

5.1 Introduction

Recommendation systems have been an effective means of providing users with recommended posts and articles based on their past viewing history. User preference-based recommendation systems enhance user engagement, but these recommendations typically hold similar views on particular issues, influencing people's beliefs and reinforcing a single viewpoint repeatedly [93]. This feature potentially traps users in "filter bubbles" [137], which can narrow users in their existing views, leading to an imbalanced understanding of certain topics. For instance, a user is deeply fascinated with the health-centric diet, believing in its numerous positive impacts. Preference-based recommendation algorithms may continuously recommend this user with favorable content about the health-centric diet while rarely presenting its drawbacks. Over time, such "reinforcement" will lead to a biased perspective, encapsulating the user within a filter bubble. This phenomenon is not exclusive to the topic of health but spans various topics, sharing the common risk of fostering biased perceptions [170], which may even lead to increased social and political polarization and extremism [130, 213]. In this case, it is important to explore an effective method for alleviating the filter bubble effects in modern recommendation systems.

Recently, some existing research works have been dedicated to examining the filter bubble phenomenon on social media platforms [93, 9, 213]. Few but growing numbers of researchers now investigate the role of AI in the development of filter bubbles, mainly concentrating

on how they could create or exacerbate filter bubble effects [214, 10, 250, 213]. These studies reveal that recommendation algorithms can negatively exacerbate filter bubble effects by reinforcing and amplifying people's beliefs, potentially causing bias [170]. Therefore, researchers started to explore novel recommendation algorithms for mitigating the filter bubble effects [73, 188]. Current solutions typically modify the underlying algorithms of recommendation systems to alleviate the filter bubble effect. However, such approaches often require substantial changes and potentially compromise user engagement and satisfaction. Considering the success and effectiveness of many existing recommendation methods [91, 231, 219, 235], an important and unsolved research question emerges, i.e., how to reduce the negative effects of filter bubbles brought on by recommendation algorithms without altering the models themselves.

To tackle the abovementioned challenges, in this paper, we propose an Agent-based Adaptive Model for Information Neutrality (AAIN), which can selectively balance the information perception and mitigate the effects of filter bubbles without altering the existing recommendation algorithms. Unlike conventional mitigation approaches that rely on algorithmic re-design, AAIN operates as an external layer, offering high modularity and seamless compatibility with existing RSs without requiring modifications to their internal architecture. We model the recommendation process of AAIN in a distributed manner utilizing Agent-based Modeling (ABM) [122]. In the ABM, the original recommendation algorithms and the system users are modeled as two agents: the Original Preference-based Agent (OPA) and the User Agent (UA). Furthermore, we embed an Adaptive Information Neutrality Agent (AINA) to mediate the interactions between the OPA and UA, aiming to selectively balance information perception and incorporate user feedback. This design ensures that AAIN can neutralize information without changing the core recommendation algorithm.

The proposed AAIN model draws inspiration from the Chinese Yin-Yang ($Y_{-}\&Y_{+}$) theory [101]. The concept of Yin-Yang emphasizes the balance and interdependence of opposing forces, suggesting that understanding both positive and negative aspects of information can lead to a more comprehensive and balanced perspective [224, 155, 194]. In the context of social networks, the application of Yin-Yang thinking becomes particularly important. The one-sided exposure can lead to polarization and the reinforcement of extreme viewpoints [93]. Encouraging users to engage with both positive and negative perspectives fosters critical thinking and helps mitigate cognitive biases. The $Y_{-}\&Y_{+}$ theory, foundational in Chinese philosophy, is symbolized by a circle divided into two halves by an 'S'-shaped curve. These two halves express the complex relationship between opposites, revealing the contrasting views on a particular subject. Y_{-} represents negativity and darkness, while Y_{+} refers to positivity and light [106]. AAIN model achieves $Y_{-}\&Y_{+}$ neutrality in recommendations,

balancing these opposing sentiment energies and providing users with broader choices rather than limiting them to specific preferences. The key contributions of this paper are summarized as follows:

- Firstly, we are the first to involve the Chinese $Y_{-}\&Y_{+}$ theory in recommendation systems. We propose a novel adaptive $Y_{-}\&Y_{+}$ Neutralization Control (YYNC) method to provide users with balanced information and broaden their exposure to diverse viewpoints.
- Secondly, we adopt a distributed modeling approach to develop a novel model, AAIN, to mitigate the effects of filter bubbles without altering the existing recommendation systems.
- Thirdly, we conduct extensive experiments on three real-world datasets to evaluate the performance of the proposed model. The experimental results explicitly show the AAIN model's capabilities in balancing recommended content, diminishing the impact of filter bubbles, and enhancing users' well-rounded understanding of the information they consume.

The remainder of this paper is organized as follows: Section 5.2 reviews the related works, including information balance, filter bubble-generated recommendation algorithms, and measurement of filter bubble effects. Section 5.3 introduces the overall framework and formal definitions. Section 5.4 elaborates on the proposed Agent-based Adaptive Information Neutrality Model. Section 5.5 demonstrates the results of the experiments and discusses the findings. Section 5.6 concludes the paper and discusses the potential directions for future research.

5.2 Related Works

In this section, related works are developed with a focus on three key areas. First, we introduce the Yin-Yang theory and its applications across various fields, highlighting its relevance to the design of recommendation systems. Second, we examine how recommendation algorithms contribute to the creation of filter bubbles through user preference modeling. Third, we discuss existing approaches to measuring and mitigating the filter bubble effect.

5.2.1 Yin-Yang Theory and Applications

The concept of Yin and Yang originated in the late primitive society of ancient China, where people observed that many opposing forces in nature, such as day and night, movement

and stillness, and life and death, were not merely contradictory but also complementary, interconnected, and interdependent [101]. This understanding gave rise to dualistic thinking and a binary classification system, forming the foundation of Yin-Yang theory [101, 224]. Initially, Yin and Yang served as a framework for explaining natural phenomena, emphasizing the mutual transformation and dynamic balance of relative forces.

As the concept evolved, it transcended natural observations and became the foundation of Chinese philosophical thought. The Yin-Yang theory shifted from interpreting isolated phenomena to viewing the world as a network of interconnected wholes [128].

Yin-Yang theory has been widely adopted in many fields. For example, in advertising, Ertz et al. applied Yin-Yang principles to enhance the credibility and effectiveness of messages by presenting balanced positive and negative information [68]. Similarly, in business intelligence, Yin-Yang-inspired models have been employed to manage dynamic systems, achieving equilibrium in complex scenarios such as portfolio management and supply-demand regulation [254]. In healthcare, Xu et al. utilized the theory for cancer treatment, conceptualizing oncogenes as Yang (promoting proliferation) and tumor suppressor genes as Yin (inhibiting proliferation), illustrating its relevance in biological systems [240].

The Chinese concept of Yin-Yang has also been applied to the realm of information, emphasizing the balance and interdependence of opposing forces. It suggests that understanding both positive and negative aspects of information fosters a more comprehensive and balanced perspective. For example, Peng and Nisbett examined cultural differences in reasoning and found that Chinese thinking is dialectical, embracing contradictions and seeking a 'middle way', which aligns with the Yin-Yang principle of balancing opposites [155]. Sundararajan discussed the coexistence of positive and negative emotions in Chinese individuals, highlighting the cultural acceptance of emotional complexity and the integration of opposing feelings, reflecting the Yin-Yang philosophy [194]. Wong further supported this perspective by demonstrating that the dynamic balance between positive and negative experiences, as conceptualized through Yin-Yang, contributes to well-being and resilience [224]. Sodan et al. provided further insights into Yin and Yang sentiments, defining Yin sentiments as softer, more passive, and introverted, while Yang sentiments are more active, aggressive, and extroverted [185]. These works emphasize that the balance between negative and positive sentiments, akin to Yin and Yang, is essential.

However, the concept of Yin-Yang has rarely been explored as a solution for addressing information imbalance problems in recommendation systems. This study contributes to this area by integrating Yin-Yang principles to balance user-specific preferences with broader sentiment diversity, providing a novel perspective on achieving more responsible recommendations. In the current setting, we focus on sentiment as the key dimension of

diversity. This is because sentiment itself can be understood through the lens of the Yin-Yang theory, where opposing emotions, such as positive opinions and negative opinions, function as complementary forces that coexist and interact within online social networks. Furthermore, many existing studies also acknowledge that sentiment is one of the important aspects of assessing diversity in a recommender system [78, 80, 230].

5.2.2 Recommendation Algorithms Leading to Filter Bubbles

In recent years, recommendation systems have been extensively studied, with most focusing on providing recommendations tailored to users' preferences [84]. However, such preference-based systems can restrict users' exposure to diverse viewpoints and perspectives, potentially contributing to the formation of filter bubbles [22].

Many existing studies have investigated the filter bubble phenomenon arising from recommender systems. For example, Vilela et al. explored the impact of filter bubbles on opinion formation using a majority-vote model, revealing that an individual's visible neighborhood can either promote or suppress polarization depending on the nature of their interactions [200]. Similarly, Li et al. examined the feedback loop in recommendation systems that leads to filter bubble formation, demonstrating how iterative narrowing of recommendations reinforces users' preexisting preferences and restricts diversity in content exposure [126].

Beyond individual-level feedback loops, Chueca analyzed the broader societal impact of filter bubbles through an agent-based model, emphasizing how algorithmic filtering and homophily-driven social networks contribute to the emergence of echo chambers and exacerbate polarization [45]. Yuan et al. further assessed the influence of AI-driven recommendation mechanisms on information filtering, finding that these systems tend to reinforce users' existing beliefs while limiting exposure to counter-narratives, thereby intensifying ideological segregation [246].

In the context of platform-specific biases, Bryant investigated preference-based recommendation algorithms on YouTube, revealing a systematic bias toward right-leaning political content and, in some cases, the amplification of extremist views [30]. Likewise, Ribeiro et al. examined YouTube's algorithmic curation, demonstrating how recommendation systems can create radicalization pathways by progressively directing users toward more extreme content over time [165]. Cinelli et al. explored the echo chamber effect on social media platforms, demonstrating how algorithmic filtering can reinforce existing beliefs by clustering users around ideologically homogeneous content [46]. Flaxman et al. analyzed online news consumption patterns and found that algorithmically curated recommendations can deepen ideological divides, particularly in politically charged environments [72].

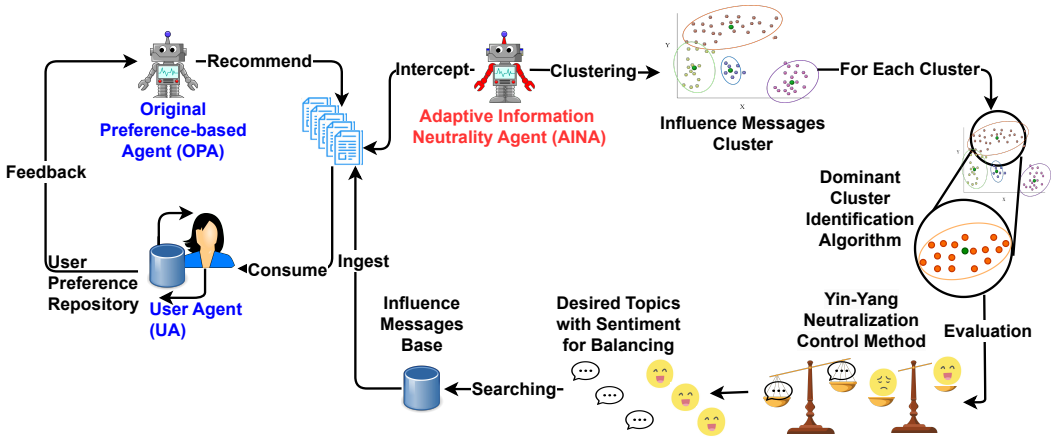


Fig. 5.1: Agent-based Framework for Adaptive Information Neutralization (AAIN).

These existing studies explicitly reveal the multifaceted role of recommendation algorithms in shaping information ecosystems. They highlight the need for mechanisms that balance personalization with content diversity to mitigate the risks of filter bubbles and echo chambers in online environments. Therefore, it becomes crucial to explore methods for quantifying their impact and developing effective mitigation strategies to ensure a more balanced and diverse information ecosystem.

5.2.3 Filter Bubble Quantification and Mitigation

Many research works have been dedicated to the analysis and quantification of the filter bubble effect. For example, Lunardi et al. proposed a metric based on homogeneity levels to quantify the extent of filter bubble formation, demonstrating that higher homogenization correlates with a greater likelihood of users being trapped within echo chambers [133]. Similarly, Bechmann and Nielbo examined the influence of recommendation algorithms on content consumption, emphasizing their role in reinforcing ideological filter bubbles by consistently prioritizing content that aligns with users' existing beliefs [18]. Expanding on this, Hu et al. investigated the ideological effects of recommendation algorithms within social networks, introducing entropy-based quantification to measure individuals' ideological isolation [93]. Their findings suggest that while recommendation systems intensify ideological segregation at an individual level, they may also facilitate occasional exposure to diverse viewpoints at a broader network level [93]. These studies collectively highlight the need for robust quantification methods to assess and address the risks posed by filter bubbles in algorithm-driven environments.

The filter bubble phenomenon is fundamentally an issue of information diversity. Many researchers aim to explore novel strategies to enhance diversity in recommendation systems. Lunardi et al. proposed a KNN item-based recommendation approach integrated with the Maximal Marginal Relevance algorithm, demonstrating that a diversified recommendation strategy can reduce homogenization and expose users to a broader range of topics [133]. Similarly, Chaitanya et al. introduced an Iterative Clustering-Based Diversity Model, which generates both similar and diverse recommendations by classifying items based on their descriptions. This approach effectively addresses several key challenges in recommender systems, including the cold start problem, filter bubbles, long-tail issues, and grey sheep challenges. By incorporating non-conversational text, which is inherently broader and more diverse, ICDM enhances content variety while maintaining contextual relevance [36]. Zhang et al. proposed a disentangled representation approach, which separates user preferences into category-independent and category-dependent components. This method enables a balanced trade-off between recommendation accuracy and diversity, ensuring that users receive both relevant and varied content [255]. Additionally, Wang et al. developed the Mixture-Channel Purpose Routing Network, which employs a purpose routing network to identify user-item interaction intents and categorize items into different channels. By utilizing purpose-specific recurrent networks, MCPRN models item dependencies within each channel, leveraging multi-objective learning to achieve an optimal balance between recommendation accuracy and diversity [217].

However, the existing diversity-enhancing techniques primarily operate at the recommendation algorithmic level, necessitating direct modifications to recommendation systems. In contrast, our proposed approach, inspired by the Yin-Yang theory, introduces an independent agent that balances opposing viewpoints without altering existing algorithms. By acting as a mediator, our AAIN fosters natural information diversity while preserving user experience and offering a more adaptive solution.

5.3 Framework and Formal Definitions

5.3.1 Overall Framework

In this research, we leverage ABM to develop the AAIN model to mitigate filter bubbles. The overall framework of the proposed AAIN is demonstrated in Figure 5.1. The AAIN model comprises three independent yet interconnected agents: OPA, UA, and AINA. Each agent employs its unique learning strategies. A user is represented as an interactive User Agent (UA) with specific preferences. The Original Preference-based Agent (OPA) functions

as the conventional preference-based recommendation system, learning the preferences of the UA and suggesting messages that align with these preferences. To counterbalance the potential bias introduced by OPA, we introduce the Adaptive Information Neutralization Agent (AINA). The role of AINA is to analyze and adjust the recommendation list from OPA by applying the Yin-Yang Neutralization Control (YYNC) method to ensure a balance of viewpoints. If the list is already neutralized, YYNC directly transfers it to the UA; otherwise, YYNC processes the list before forwarding it to the UA. Through this mechanism, AINA integrates complementary viewpoints into OPA's recommendations, offering UA a balanced information experience. This approach aims to harmonize the viewpoints within the recommendations, ensuring diversified and unbiased information is provided to the user.

5.3.2 Formal Definitions

Definition 1: User Agent (v_i). User Agent refers to an individual within a recommendation context whose representation is captured through an embedding process. At any given time t, the embedding of v_i, i.e., $\mathbf{v_i}$, is obtained by aggregating the embeddings of all messages interacted with v_i, utilizing the Hadamard product [138] for combination. UA receives recommendations from AINA and decides whether to accept or reject these recommendations based on their preferences.

Definition 2: Influence Messages ($M^t_{v_i}$). Influence messages in recommendation systems generally refer to the communications that affect users' opinions, behaviors, or decisions through their content and dissemination. They are the concrete presentation of social influence. In the proposed model, the OPA recommends a set of messages at time t based on the preference of v_i, denoted as $M^t_{v_i} = \{msg_1,...,msg_j\}$. Message msg_x can be represented as a tuple, i.e., $(q(msg_x), T(msg_x), o(msg_x))$, having the message's content text $q(msg_x)$, the associated topic $T(msg_x)$, and the sentiment intensity $o(msg_x) \in [0,1]$. In this paper, the $Y_-\&Y_+$ neutralization refers to the sentiment balance for specific topics in $M^t_{v_i}$.

Definition 3: Yin and Yang ($Y_-\&Y_+$). Yin-Yang illustrates the dual nature of interactions and sentiments that shape users' experiences and perceptions. Messages are classified based on their sentiment value, $o(msg_x)$. If $o(msg_x)$ ranges from 0 to 0.5, it expresses the negative sentiment and is classified as Y_-. Values exceeding 0.5 indicate positive sentiment and are classified as Y_+. A value of precisely 0.5 denotes a neutral sentiment. The intensity of the sentiment is considered more extreme as $o(msg_x)$ diverges further from 0.5, whether it falls under Y_- or Y_+.

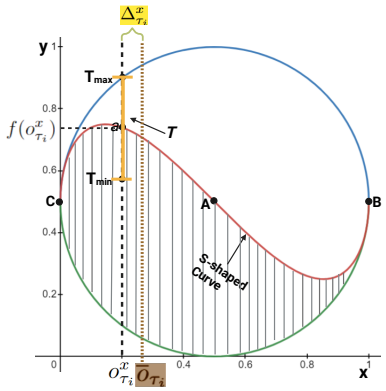


Fig. 5.2: The Yin-Yang Model for Information Neutralization

5.3.3 Yin-Yang Neutralization

Yin-Yang Neutralization, in this context, refers to balancing positive and negative interactions to foster a harmonious social environment and mitigate polarizing effects. Figure 5.2 depicts the proposed information neutralization model, inspired from the traditional Chinese philosophy of Yin-Yang.

This diagram features a circle centered at (0.5, 0.5) with a radius of 0.5, divided by an 'S'-shaped curve (segment **C**-**A**-**B**) that represents the transition between Y_- (below) and Y_+ (above) sentiments. This curve embodies the ideal $Y_-\&Y_+$ neutralization [106]. The x and y axes represent the sentiment scores of messages for a specific topic, ranging from 0 to 1. For any given message with sentiment score x, the corresponding y on the 'S'-shaped curve represents the complementary sentiment required to neutralize this message. Inspired by [40], the mathematical representation of ideal $Y_-\&Y_+$ neutralization is formulated as follows:

$$f(o^x_{\tau_i}) = \frac{1}{2} + \left(1 - 2o^x_{\tau_i}\right)\sqrt{\frac{1}{4} - \left(o^x_{\tau_i} - \frac{1}{2}\right)^2} \tag{5.1}$$

Here, $o^x_{\tau_i} \in [0,1]$ is an individual sentiment, which is the sentiment of a message under topic τ_i. $o^x_{\tau_i}$ is a sentiment score within a collection $O_{\tau_i} = \{o^1_{\tau_i}, ..., o^x_{\tau_i}\}$. The collection O_{τ_i} encompasses various sentiment scores associated with a specific topic τ_i in a message set $M^t_{v_i}$. A point on the 'S'-shaped curve represents a $Y_-\&Y_+$ neutralization pair.

The vertical line of $\overline{o}_{\tau_i}$ represents the public sentiment of all users towards a topic τ_i. It is calculated as the average sentiment of all users related to that topic, which is defined in Equation 5.2. $\Delta^x_{\tau_i} = |\overline{o}_{\tau_i} - o^x_{\tau_i}|$ is the distance between public sentiment and individual sentiment, which is the absolute value of these two values.

$$\overline{o}_{\tau_i} = \frac{\sum\limits_{msg_x \in \hat{M}_{v_j,\tau_i} | v_j \in V} o^x_{\tau_i}}{\sum\limits_{v_j \in V} |\hat{M}_{v_j,\tau_i}|} \tag{5.2}$$

The aim of $Y_-\&Y_+$ neutralization is to achieve the sentiment balance in O_{τ_i} for the designated topic τ_i. A tolerance interval $\boldsymbol{T}$ (like segment $\mathbf{T_{max}}$-$\mathbf{T_{min}}$ in Figure 5.2) is established for accepted $Y_-\&Y_+$ neutralization. The rationale for establishing this interval is that users with extreme preferences tend to have a narrower acceptance threshold, whereas those with more moderate views possess a broader acceptance range. When $x = o^x_{\tau_i}$, $\mathbf{T_{max}}$ can be expressed as $\frac{1}{2} + \sqrt{\frac{1}{4} - \left(o^x_{\tau_i} - \frac{1}{2}\right)^2}$. The sentiment score within O_{τ_i} that is closest to the 'S'-shaped curve and lies in interval $\boldsymbol{T}$ is considered a $Y_-\&Y_+$ neutralization pair with $o^x_{\tau_i}$, denoted as $(o^x_{\tau_i}, o^y_{\tau_i})$. The value $o^y_{\tau_i}$ is the complementary sentiment to $o^x_{\tau_i}$ and equals $f(o^x_{\tau_i})$ if point $(o^x_{\tau_i}, o^y_{\tau_i})$ lies on the 'S'-shaped curve. Therefore, $o^y_{\tau_i}$ should belong to the interval $\left[2f(o^x_{\tau_i}) - \mathbf{T_{max}}, \mathbf{T_{max}}\right]$. If all scores in O_{τ_i} can be matched accordingly, this collection is considered to have reached $Y_-\&Y_+$ neutralization.

5.4 Agent-Based Adaptive Information Neutrality Model

In this section, we detail the operations of the AAIN model, which encompasses three interrelated yet independent agents: Original Preference-based Agent (OPA), Adaptive Information Neutrality Agent (AINA), and User Agent (UA). Specifically, OPA recommends influence messages to the UA based on user preferences. AINA selectively intercepts these messages and outputs a final $Y_-\&Y_+$ neutralized recommendation list to the UA. UA responds to outputs from AINA based on current user preferences and archives these responses into the user preference repository. Finally, OPA receives the responses from UA and provides the next set of recommendations.

5.4.1 Original Preference-based Agent

The OPA acts as a fundamental preference-based recommendation system, utilizing algorithms such as Collaborative Filtering (CF) [204] to curate influence messages. It collaborates with UA to tailor these recommended messages based on UA's preferences. Subsequently, AINA processes the selected messages to neutralize information, generating the final recommendation list.

5.4.2 Adaptive Information Neutrality Agent

The AINA plays an important role in the AAIN model by facilitating information neutralization. After receiving a recommendation list from OPA, AINA clusters the messages. We introduce the Dominant Cluster Identification Algorithm (DCIA) to pinpoint the cluster for $Y_{-}\&Y_{+}$ neutralization. Next, the Yin-Yang Neutralization Control (YYNC) method analyzes and balances the information within the recommendations. AINA selects messages based on the $Y_{-}\&Y_{+}$ outcomes from YYNC, enhances OPA's recommendations with these selections, and sends the refined list to UA.

Message Clustering

Message clustering is an initial step in AINA, setting the stage for subsequent information neutralization. This step involves grouping messages based on semantic and thematic similarities, which are quantified using both textual embeddings and topic representations. We employ BERT [108] for generating textual embeddings for each message (msg_x). Each message is tagged with a topic $T(msg_x)$.

Textual Similarity: Cosine distance [20] is leveraged to estimate the similarity of two message embeddings $\mathbf{msg_i}$ and $\mathbf{msg_j}$:

$$\rho(msg_i, msg_j) = \frac{\mathbf{msg_i} \cdot \mathbf{msg_j}}{\|\mathbf{msg_i}\| \left\|\mathbf{msg_j}\right\|} \tag{5.3}$$

where $\mathbf{msg_i}$ and $\mathbf{msg_j}$ denote textual embeddings, $\mathbf{msg_i} \cdot \mathbf{msg_j}$ refers to the dot product of the textual embeddings, and $\|\mathbf{msg_i}\|$ and $\left\|\mathbf{msg_j}\right\|$ denote the corresponding Euclidean norms.

Topic Similarity: $T(msg_x)$ denotes the singular topic assigned to message msg_x. As each message is only associated with one topic, a single topic is related to multiple messages. We can obtain the topic embedding $\mathbf{T_x}$ by aggregating related messages' embeddings. The similarity of messages with topic labels is quantified using cosine distance:

$$\rho(T(msg_i), T(msg_j)) = \frac{\mathbf{T_i} \cdot \mathbf{T_j}}{\|\mathbf{T_i}\| \left\|\mathbf{T_j}\right\|}, \tag{5.4}$$

Unified Distance Metric: A unified distance metric considers both textual and topical similarity, which is formulated as follows:

$$\begin{aligned} d(msg_i, msg_j) =& \alpha \times \rho(msg_i, msg_j) \\ &+ (1-\alpha) \times \rho(T(msg_i), T(msg_j)), \end{aligned} \tag{5.5}$$

where α balances the influence of the textual and topical embeddings in clustering. In the current setting, we set α as 0.5 to treat both equally.

Clustering Algorithm: We use k-means to partition messages into k clusters, optimized by a combined distance metric $d(msg_i, msg_j)$. The centroids μ_k are re-calibrated by computing the mean of all messages designated to cluster C_k:

$$\mu_k = \frac{1}{|C_k|} \sum_{msg_i \in C_k} msg_i \tag{5.6}$$

This clustering prepares the ground for the ensuing $Y_- \& Y_+$ analysis aimed at neutralizing sentiment imbalance.

Dominant Cluster Identification Algorithm

DCIA is designed to facilitate $Y_- \& Y_+$ neutralization by concentrating on key clusters, prioritizing those that are larger and more stable based on their size and entropy. A critical feature of the DCIA is the Memory Mechanism, which adapts its strategies based on the feedback from UA derived from their interactions. If UA consistently rejects the recommendation, DCIA shifts the focus of $Y_- \& Y_+$ neutralization to a different cluster.

Entropy Calculation for Each Cluster: Entropy measures the diversity within a cluster, specifically focusing on the variety of topics and sentiments. For a given cluster C_k and topic τ_i, the average sentiment score P' is estimated as:

$$P'(\tau_i, C_k) = \frac{\sum_{j \in C_k} o^j_{\tau_i}}{\sum_{j \in C_k}}, \tag{5.7}$$

where $o^j_{\tau_i}$ denotes the sentiment score of message msg_j towards topic τ_i. When $o^j_{\tau_i} \in [0,1]$, the range of $P'(\tau_i, C_k)$ also falls within $[0,1]$.

The entropy of cluster C_k is defined as:

$$H(C_k) = -\sum_{\tau_i \in T} P'(\tau_i, C_k) \log_2 P'(\tau_i, C_k) \tag{5.8}$$

In Equation 5.8, $H(C_k)$ measures the diversity of sentiment associated with the topics within the cluster. A higher value signifies a more diverse array of viewpoints within the cluster.

Cluster Importance Evaluation Algorithm: The importance of a cluster ($Imp(C_k)$) is determined by its size (S_k) and entropy ($H(C_k)$). We use normalized values for size S'_k and entropy $H'(C_k)$ to calculate this. Thus,

$$S'_k = \frac{S_k}{\sum_{j=1}^{n} S_j}, \tag{5.9}$$

where n denotes the total number of clusters. Furthermore, the entropy values are normalized as follows:

$$H'(C_k) = \frac{H(C_k)}{\max_{j=1,\ldots,n} H(C_j)} \tag{5.10}$$

Finally, $Imp(C_k)$ is defined as:

$$Imp(C_k) = \frac{2 \cdot (S'_k \times H'(C_k))}{(S'_k + H'(C_k))} \tag{5.11}$$

Memory Mechanism The Memory Mechanism guides the $Y_{-}\&Y_{+}$ neutralization efforts toward topics within a specific cluster, informed by UA's feedback. A threshold, denoted by **R**, is set to limit the allowable number of rejections for messages produced by AINA within the final recommendation list. If the number of rejections by a user v_i surpasses **R**, AINA adjusts the $Y_{-}\&Y_{+}$ neutralization target to the subsequent cluster of importance.

Yin-Yang Neutralization Control Method

The primary task of the Yin-Yang Neutralization Control (YYNC) Method is to assess the $Y_{-}\&Y_{+}$ balance of topics within a specific cluster and apply balancing measures to those topics that are found to be imbalanced.

In some cases, a topic may naturally evoke overwhelmingly one-sided sentiment, for example, a product that receives predominantly negative reviews from most customers. We believe such extreme sentiments accurately reflect the actual situation and do not require neutralization. Therefore, the distance $\Delta^x_{\tau_i}$ between public sentiment ($\overline{o}_{\tau_i}$) and individual sentiment ($o^x_{\tau_i}$) is considered to filter out such topics before neutralization (refer to in Figure 5.2).

A topic filtering threshold θ is used to determine whether a message should undergo Yin-Yang Neutralization Control (YYNC). If the sentiment distance $\Delta^x_{\tau_i} \leq \theta$, the topic is treated as naturally "imbalanced" topic, and the message msg_x is considered to be in line with the public sentiment of topic τ_i, therefore excluded from the YYNC process and does not require intervention. However, if $\Delta^x_{\tau_i} > \theta$, the individual sentiment differs significantly from the public sentiment, and YYNC is applied to perform neutralization on msg_x.

Once YYNC receives the sentiment score collection O_{τ_i}, it first determines whether O_{τ_i} meets the $Y_{-}\&Y_{+}$ neutralization criteria. Sentiment scores within O_{τ_i} are sorted by intensity, placing those farthest from the neutral value of 0.5 at the top of the list. This sorted list is then divided into two halves; the first half, $O^1_{\tau_i}$, contains scores designated as x values in Figure 5.2, while the second half, $O^2_{\tau_i}$, comprises potential matching y values for neutralization. Scores in $O^1_{\tau_i}$, being further from 0.5, are targeted for neutralization due to their more intense sentiments.

If the total number of scores in O_{τ_i} is odd, the most neutral sentiment score, typically the last one in the sorted list, is removed to balance the halves. If each score in $O^1_{\tau_i}$ can be successfully paired with a score from $O^2_{\tau_i}$ according to the $Y_{-}\&Y_{+}$ neutralization criteria, AINA will not intervene with OPA's operations but will directly pass the information to UA. If pairing fails, indicating an imbalance in O_{τ_i}, the YYNC initiates the neutralization process. In cases of imbalance, YYNC employs Algorithm 1 to find the most suitable sentiment score from the database to create $Y_{-}\&Y_{+}$ pairs, thereby achieving sentiment balance.

Searching

The final phase of the $Y_{-}\&Y_{+}$ neutralization process, referred to as "Searching", revolves around treating the sentiment list (*complement_list*) generated by the YYNC method as a search query. This query is subsequently utilized to retrieve relevant messages from the information message base. These retrieved messages are then finally merged with the recommendation ($M^t_{v_i}$) provided by OPA, culminating in a unified message set that is presented as a recommendation to UA.

Illustrated in Figure 5.3, the outputs from the YYNC method function as the query for the "Searching" step. This query is input into the Information Message Base, initiating a search within the base to extract messages with matching topics and corresponding sentiment scores. In the depicted example of Figure 5.3, assuming the sentiment query value associated with topic τ_i is 0.035, we search all messages of topic τ_i with 0.035 sentiment score from the information message base. Subsequently, a pre-trained embedding model is engaged to extract features from each message. A similarity assessment is executed between the feature of each message and user preferences $\mathbf{v_i}^t$ to ensure the output message exhibits the highest similarity with user preferences. Essentially, for each sentiment query, a corresponding message output is generated. These outputs, when combined with $M^t_{v_i}$ from OPA, form AINA's final output, denoted as $M^t_{v_iR}$.

In scenarios where an influence message receives overwhelmingly negative reviews, the effectiveness of the balancing process may be limited if the influence message base contains few or no counterbalancing messages. However, discussions on most topics generally include

Algorithm 2 Yin-Yang Neutralization Control Method.

Input:
- Sentiment data pool for topic τ_i: $O_{\tau_i}^{pool} = \{o_{\tau_i}^i, ..., o_{\tau_i}^n\}$
- $Y_-\&Y_+$ neutralization target for topic τ_i: $O_{\tau_i} = \{o_{\tau_i}^j, ..., o_{\tau_i}^m\}$
- tolerant: tol
- topic filtering threshold θ

Output: A complementary list for $Y_-\&Y_+$ neutralization: *complement_list*

function COMPAREPUBLICSENTIMENT(τ_i, $M_{v_j}^t$)
    Create an empty set $M_{v_j}^{t'}$
    Calculate public sentiment towards topic τ_i as $\overline{o}_{\tau_i}$
    Use Equation 5.2 to compute $\overline{o}_{\tau_i}$
    for each message msg_x in $M_{v_j}^t$ **do**
        Calculate individual sentiment $o_{\tau_i}^x$
        if $\left|\overline{o}_{\tau_i} - o_{\tau_i}^x\right| < \theta$ **then**
            continue
        else
            $M_{v_j}^{t'} = M_{v_j}^{t'} \cup \{msg_x\}$
        end if
    end for
     return The filtered message list $M_{v_j}^{t'}$ for Yin-Yang Neutralization
end function
function IFBALANCE(O_{τ_i})
    Sort O_{τ_i} by closeness to neutral sentiment 0.5
    $O_{\tau_i}^1, O_{\tau_i}^2 \leftarrow$ split_list(O_{τ_i})
    Initialize *remain_list* to track unpaired sentiments
    for $o_{\tau_i}^x$ **in** $O_{\tau_i}^1$ **do**
        $min, max \leftarrow$ get_tolerance$(o_{\tau_i}^x)$ ▷ The distance between ($o_{\tau_i}^x$, *min*) and ($o_{\tau_i}^x$, *max*) is the tolerance interval for $Y_-\&Y_+$ neutralization
        $o_{\tau_i}^y \leftarrow$ find_complement_num$(o_{\tau_i}^x, O_{\tau_i}^2, min, max)$
        if $o_{\tau_i}^y$ is in $O_{\tau_i}^2$ **then**
            $O_{\tau_i}^2$.remove$(o_{\tau_i}^y)$
        else
            remain_list.append$(o_{\tau_i}^x)$
        end if
    end for
    if $O_{\tau_i}^2$ is None **then**
        O_{τ_i} is naturalization
    else
        return *remain_list*
    end if
end function
function FINDMATCHSCORE(*remain_list*, $O_{\tau_i}^{pool}$)
    Initialize *complement_list* to include sentiments added for balance
    for $o_{\tau_i}^x$ **in** *remain_list* **do**
        $min, max \leftarrow$ get_tolerance$(o_{\tau_i}^x)$
        $o_{\tau_i}^y \leftarrow$ find_complement_num$(O_{\tau_i}^{pool}, min, max)$
        if $o_{\tau_i}^y \neq$ None **then**
            complement_list.append$(o_{\tau_i}^y)$
            return *complement_list*
        else

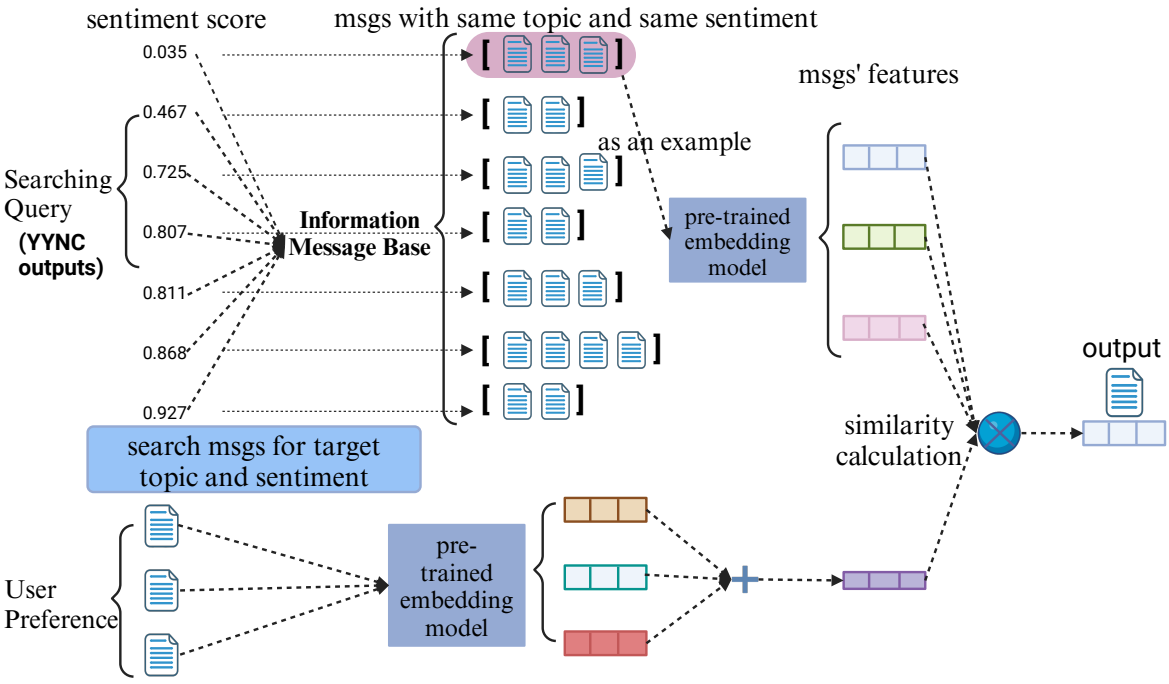


Fig. 5.3: An example of the searching process in AINA

both positive and negative perspectives. Therefore, in this study, we assume that the influence message base is sufficiently large and diverse to provide adequate counterbalancing messages.

5.4.3 User Agent

The UA receives a $Y_- \& Y_+$ neutralized recommendation list from AINA. It responds to this list based on user preferences from historical behaviors and decides whether to accept or reject the messages from AINA. To simulate the user's responses to AINA's recommendations, an offline Acceptance Probability Algorithm is employed. This algorithm assesses the similarity between preferences of UA and the messages in $M^t_{v_iR}$ using the cosine similarity:

$$\rho(v_i^t, M^t_{v_iR}) = \frac{\mathbf{v_i}^t \cdot \mathbf{M}^t_{v_iR}}{||\mathbf{v_i}^t|| ||\mathbf{M}^t_{v_iR}||}, \tag{5.12}$$

where $\mathbf{M}^t_{v_iR}$ represents the integrated embedding of all messages in $M^t_{v_iR}$, and $\mathbf{v_i}^t$ signifies v_i's preferences at time t.

This algorithm incorporates a random value to model v_i's responses. If the cosine similarity $\rho(v_i^t, M_{v_iR}^t)$ exceeds this random threshold, it indicates that v_i accepts the recommended message. Otherwise, the message is rejected. Based on these acceptance or rejection decisions, feedback from UA is forwarded to OPA to generate next-round messages. Simultaneously, this feedback updates the user preference repository, enhancing the accuracy of future recommendations from AINA.

5.5 Experiments and Analysis

In this section, we conduct extensive experiments to evaluate the proposed AAIN model in terms of fostering sentiment diversity, realizing $Y_-\&Y_+$ neutralization, and consequently mitigating filter bubbles.

5.5.1 Datasets and Settings

We utilize three real-world datasets, i.e., the Microsoft News dataset (MIND) [1], IMDB [2] dataset, and Goodreads Book Datasets (GBD) [3]. MIND is a public news recommendation dataset with user interaction data from Microsoft News. MIND comprises data from 5,000 users, including 230,117 user reading records and 51,287 news. IMDB is a movie recommendation dataset. We use the rating records provided by the dataset. It comprises 25,000 movie rating records from 333 users and a range of 2,586 movies. The GBD dataset specializes in books, featuring information from 1,542 users and 31,740 books. The book dataset includes details such as book IDs, categories, titles, abstracts, and more. We set the topic filter threshold $\theta = 0.1$ because topics in these datasets do not present extreme sentiment polarization.

Given the inherent impracticality and significant cost of online testing for researchers, we have developed an offline simulation approach to evaluate the proposed method. The Acceptance Probability Algorithm described in Section 5.4.3 simulates user feedback for recommendations.

5.5.2 Evaluation Metrics

Three evaluation metrics have been adopted for the experiments.

[1] https://msnews.github.io/

[2] https://www.kaggle.com/datasets/meastanmay/imdb-dataset?select=tmdb_5000_movies.csv/

[3] https://www.kaggle.com/datasets/bahramjannesarr/goodreads-book-datasets-10m/

- **Diversity**: Measures the range of sentiments across recommendations, indicating a system's ability to offer varied perspectives. It consists of sentiment coverage and repetition rate (RR), where coverage refers to the sum of sentiments for a particular topic in the recommendation list divided by the sum of sentiments for that topic in the data pool and then averaged. The RR calculates the average sentiment redundancy for a specific topic, as the topic is τ_i, $RR = \frac{\sum_{i=1}^{n} RR_{\tau_i}}{n}$.

- **Accuracy**: Reflects how recommendations align with user preferences. It's evaluated through Hit (1 if at least one message is accepted in a recommendation, 0 otherwise) and Precision (Pre), the ratio of accepted items in the recommendation.

- **Yin-Yang Neutralization Degree (best_diff)**: Quantifies the degree of $Y_{-}\&Y_{+}$ neutralization in recommendation systems, we introduce a metric defined by the formula: $\left|\sum_{x=1}^{n} f(o_{\tau_i}^{x}) - \sum_{y=1}^{m} o_{\tau_i}^{y}\right|$, where f represents the curve function detailed in Equation 5.1. Here, $o_{\tau_i}^{x}$ and $o_{\tau_i}^{y}$ denote the individual sentiments within the list O_{τ_i}, corresponding to Y_{-} and Y_{+} respectively. The variables m and n are the counts of Y_{-} and Y_{+} scores in the list. A "best_diff" value of 0 signifies that the list O_{τ_i} has achieved the perfect neutralization.

5.5.3 Baselines

We assess the performance of AAIN in comparison with several established baseline methods as follows:

- **NGCF** [221]: Enhances recommendation by modeling collaborative signals with user-item graphs.

- **LGCN** [91]: Focuses on neighborhood aggregation for efficient training; outperforms NGCF.

- **DGCF** [222]: Disentangles user-item relationships to capture user intent diversity.

- **ENMF** [38]: Employs three new optimization methods in a neural matrix factorization framework for better performance.

- **SGL** [231]: Incorporates self-supervised learning into traditional GCNs for improved accuracy and robustness.

- **DiffRec**[219]: Leverages diffusion-based graph structures to propagate user preferences and item features more effectively, capturing both local and global collaborative signals for enhanced recommendation quality.

- **LDiffRec**[219]: An extension of DiffRec that introduces layer-wise diffusion mechanisms, improving the capture of multi-scale user preferences and item relationships for better recommendation accuracy and interpretability.

For each baseline, we also explore its integration with the AAIN model, denoted as "Method$_{AAIN}$" (e.g., SGL$_{AAIN}$), producing 14 comparative baselines to assess the effectiveness of AAIN on mitigating filter bubbles and achieving neutralization goals.

5.5.4 Experiment 1: Evaluation of Diversity and Accuracy

The first experiment evaluates the AAIN model's impact on enhancing diversity and accuracy within preference-based recommendation systems. By leveraging metrics such as Coverage and RR as metrics for diversity, as well as Hit and Pre for accuracy, our objective is to show how the AAIN model enhances sentiment diversity without significantly compromising recommendation precision. The experiment conducted on the MIND, IMDB, and GBD datasets involved 10 recommendation models applied to 5 randomly selected users from each dataset (U45, U46, U103, U10022, U23393 from MIND, U22, U62, U63, U88, U109 from IMDB, and U10013, U10199, U9188, U7241, U7607 from GBD), illustrating the benefits of incorporating the AAIN model in overcoming filter bubbles and fostering a neutral recommendation environment. The evaluation results are demonstrated in Table 5.1, where "$\Delta(\%)$" denotes the percentage change due to AAIN.

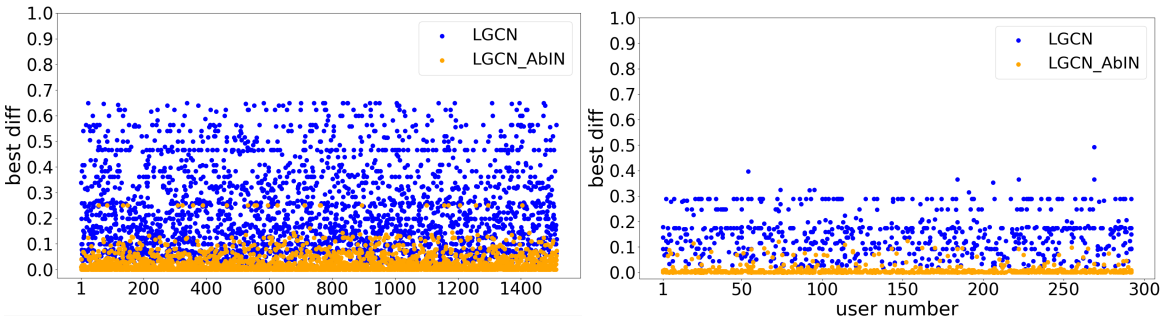
The results demonstrate that incorporating the AAIN model significantly enhances sentiment coverage across three datasets from different domains while reducing sentiment redundancy. This leads to a richer user experience by offering a broader range of sentiments and further validates the model's generalizability. Despite mixed outcomes in precision, the findings underscore the AAIN model's ability to balance diversity and accuracy effectively. Overall, this experiment underscores the AAIN model's role in enhancing the diversity of preference-based recommendation systems. By broadening the variety of content recommended to users, the model addresses critical challenges such as filter bubbles.

5.5.5 Experiment 2: Neutralization Evaluation

This experiment evaluates the AAIN model's capability to enhance neutralization in recommendation systems, specifically aiming to balance Y_- and Y_+ to provide a more neutral recommendation environment and reduce the impact of filter bubbles. The "best_diff" matrix serves as a quantitative measure of neutralization extent. This study includes two datasets: 2,000 users were randomly selected from MIND, and all users from IMDB were included.

Table 5.1: Comparative Analysis of Recommendation Models with and without the AAIN Enhancement

	MIND				IMDB				GBD			
	Diversity		Accuracy		Diversity		Accuracy		Diversity		Accuracy	
Models	Coverage	RR	Pre	Hit	Coverage	RR	Pre	Hit	Coverage	RR	Pre	Hit
SGL	0.0009	0.0349	0.4920	1	0.0136	0.0031	0.4619	1	0.0013	0.0162	0.496	1
SGL_{AAIN}	0.0012	0.0169	0.5096	1	0.0124	0.0040	0.5052	1	0.0013	0.0116	0.4794	1
Δ(%)	↑ 28.33	↓ 51.56	↑ 3.57	-	↓ 8.82	↑29.03	↑ 9.38	-	-	↓ 28.58	↓ 3.35	-
NGCF	0.0011	0.0399	0.4660	1	0.0135	0.0030	0.4880	1	0.0013	0.0130	0.448	1
$NGCF_{AAIN}$	0.0012	0.0282	0.4656	1	0.0144	0.0057	0.5145	1	0.0014	0.0129	0.4655	1
Δ(%)	↑ 8.36	↓ 29.44	↓ 0.09	-	↑ 6.78	↑90.00	↑ 5.43	-	↑ 7.69	↓ 1.51	↑ 3.91	-
LGCN	0.0010	0.0413	0.538	1	0.0116	0.0122	0.516	1	0.0013	0.0166	0.498	1
$LGCN_{AAIN}$	0.0012	0.0249	0.5103	1	0.0129	0.0041	0.5366	1	0.1489	0.1000	0.4890	1
Δ(%)	↑ 23.50	↓ 39.62	↓ 5.15	-	↑ 11.16	↓ 66.75	↑ 3.99	-	↑ 11.70	↓ 40.03	↑ 1.81	-
DGCF	0.0012	0.0471	0.49	1	0.0143	0.0044	0.53	1	0.0013	0.0051	0.468	1
$DGCF_{AAIN}$	0.0012	0.0267	0.5112	1	0.0140	0.0084	0.5247	1	0.0016	0.0062	0.5152	1
Δ(%)	-	↓43.21	↓4.44	-	↓2.10	↑90.91	↓ 1.00	-	↑ 18.93	↑23.75	↑ 10.09	-
DiffRec	0.0010	0.04809	0.492	1	0.0157	0.0172	0.476	1	0.0012	0.0160	0.49	1
$DiffRec_{AAIN}$	0.0011	0.0221	0.5010	1	0.0158	0.0148	0.4677	1	0.0013	0.0128	0.5046	1
Δ(%)	↓ 8.10	↓54.02	↑1.82	-	↑13.56	↓83.35	↓ 3.46	-	↑8.48	↓40.49	↑ 1.27	-
LDiffRec	0.0011	0.0350	0.478	1	0.0146	0.0118	0.496	1	0.0010	0.0135	0.552	1
$LDiffRec_{AAIN}$	0.0012	0.0249	0.5103	1	0.01678	0.0053	0.5156	1	0.0013	0.0128	0.5046	1
Δ(%)	↑ 6.93	↓ 28.92	↑ 6.76	-	↑ 14.86	↓ 54.63	↑ 3.96	-	↑ 19.00	↓ 5.49	↓ 8.59	-
ENMF	0.0010	0.0333	0.490	1	0.0138	0.0025	0.512	1	0.0013	0.0340	0.4861	1
$ENMF_{AAIN}$	0.0011	0.0304	0.5237	1	0.0174	0.0127	0.5531	1	0.0017	0.0136	0.5301	1
Δ(%)	↑7.80	↓ 8.78	↑ 6.87	-	↑ 26.12	↓ 408.92	↑ 8.04	-	↑ 26.25	↓ 60.2	↑ 9.04	-



(a) (b)

Fig. 5.4: An analysis of the efficacy of the AAIN model in the context of a single-instance Yin-Yang neutralization task on two datasets. (a) Yin-Yang neutralization evaluation on MIND dataset. (b) Yin-Yang neutralization evaluation on IMDB dataset.

In the scatter plots referenced (Figure 5.4), the x-axis indicates user numbers, and the y-axis represents the degree of $Y_{-}\&Y_{+}$ neutralization (best_diff) across recommended topics for each user in a single recommendation. The figures reveal that the AAIN-enhanced model ($LGCN_{AAIN}$), depicted by yellow dots, significantly improves $Y_{-}\&Y_{+}$ neutralization compared to the LGCN model (blue dots), with many instances nearing perfect neutralization

Table 5.2: Comparison of LGCN and LGCN$_{AAIN}$ on MIND and IMDB datasets

	MIND				**IMDB**			
Cluster Size	Time(m)	LGCN	LGCN$_{AAIN}$	Improve(%)	Time(m)	LGCN	LGCN$_{AAIN}$	Improve(%)
1	76.58	0.2332	0.0244	89.54	0.598	0.7287	0.0165	97.72
2	38.56	0.2358	0.0203	91.38	0.454	0.6622	0.0110	**98.32**
3	34.06	0.2384	0.0192	91.97	0.400	0.6650	0.0111	98.31
4	30.41	0.2442	0.0191	92.18	0.355	0.8391	0.0264	96.84
5	30.30	0.2333	0.0180	92.27	0.358	0.9169	0.0390	95.74
6	29.18	0.2293	0.0174	**92.32**	0.364	0.9388	0.0431	95.40
7	28.86	0.2263	0.0230	89.85	0.370	0.9750	0.0464	95.23
8	26.34	0.2299	0.0225	90.20	0.379	0.9697	0.0503	94.80

(approaching 0). This suggests that integrating AAIN can assist in achieving a more neutral and balanced recommendation environment.

The experimental results underscore the AAIN model's impact on improving $Y_{-}\&Y_{+}$ neutralization within recommendation systems. By consistently lowering the "best_diff" values across a broad user base, the AAIN-enhanced model demonstrates its superiority over preference-based approaches.

5.5.6 Experiment 3: Impact of Cluster Sizes

In this parameter experiment, we assess the impact of varying cluster sizes on the speed and extent of $Y_{-}\&Y_{+}$ neutralization by the recommendation system, which is configured to suggest 10 messages per recommendation. This setup remains consistent with that of Experiment 2, except for the cluster sizes examined.

Table 5.2 demonstrates an acceleration in the neutralization process as the cluster size increases, as noted in the "time" column. This column quantifies the degree of $Y_{-}\&Y_{+}$ neutralization (reflected by "best_diff") achieved by the AAIN-enhanced model compared to the LGCN model. Despite minor fluctuations, cluster size 6 is selected for the MIND dataset for optimal performance. Conversely, the IMDB dataset exhibits a significant shift in neutralization performance. The "Improve(%)" diminishes from 98.32% to 94.80% as cluster size increases and stabilizes from cluster size 4 onwards. This suggests that the messages within the targeted cluster appear stable beyond this point. Cluster size 2 is chosen for the IMDB dataset, demonstrating the best performance on $Y_{-}\&Y_{+}$ neutralization.

5.5.7 Discussions

The above three experiments offer an extensive evaluation of the effectiveness of the proposed AAIN model. They also demonstrate the AAIN model's multifaceted capabilities, including

its performance on recommendation diversity, precision, and $Y_- \& Y_+$ neutralization. The following insights can be discovered from the experiments.

- The AAIN model shows a strong capability to enhance recommendation diversity without severely influencing the accuracy of recommendations. This suggests that AAIN can effectively manage the trade-off between broadening diversity and maintaining user-specific relevance.
- The neutralization of sentiments (Y_- and Y_+) by the AAIN model indicates that the model can significantly mitigate biases inherent in preference-based recommendation systems. This neutralization is crucial for promoting a more balanced information exposure, which could lead to a more diverse user perspective.
- Selecting appropriate cluster sizes turns out to be a major consideration in the speed and degree of neutralization. Specifically, different datasets may require different configurations, e.g., cluster sizes, to achieve the best performance.

5.6 Conclusion and Future Work

This paper presented the Agent-based Adaptive Information Neutrality (AAIN) model, a novel approach inspired by the traditional Chinese Yin-Yang theory, designed to mitigate filter bubble effects in recommendation systems without changing existing algorithms. Unlike conventional methods that focus on modifying the underlying mechanics of recommendation algorithms, the AAIN model introduces a multi-agent framework that integrates with the existing preference-based recommendation systems to enhance neutrality and diversity.

The key strength of the AAIN model lies in its ability to balance contrasting information perspectives ($Y_- \& Y_+$), thus promoting a more neutral understanding of diverse topics. This balance is achieved through the innovative YYNC mechanism, which ensures that users are exposed to neutralizing views, counteracting the bias caused by preference-based algorithms. Additionally, the AAIN model includes an adaptive mechanism that activates the $Y_- \& Y_+$ neutralization strategy, specifically in situations where information is imbalanced. The extensive experiments conducted in this research demonstrate the AAIN model's superior capability to mitigate the impact of filter bubbles and enhance the users' engagement with a broader spectrum of information. These results sufficiently validate the model's effectiveness in information neutralization.

Nevertheless, due to the simplified Yin-Yang theory incorporated in this work, we will explore and utilize the Yin-Yang theory more deeply to mitigate information imbalance in the

future. Meanwhile, we plan to investigate the potential for user-driven customization options within the AAIN model. This could provide insights into user preferences for information diversity. Such an approach could allow users to adjust the level of information neutrality or diversity they prefer, potentially enhancing user satisfaction and engagement. Furthermore, we aim to consider other aspects of diversity that could be suitable for Yin-Yang theory in future work, broadening its applicability and impact in addressing information balance.

Chapter 6

Conclusion

6.1 Introduction

This thesis advances the field of RSs by addressing several critical challenges, including accuracy, diversity, and fairness. The motivation behind this research was to overcome key limitations of traditional and ML-based RSs, such as "filter bubbles", data sparsity, cold start issues, and the ethical implications these systems can create. While traditional and ML-based RSs have demonstrated success in personalizing content, they often amplify bias and limit information diversity, which can negatively impact user trust and experience.

By introducing new models that push the boundaries of RSs, this thesis presents solutions that ensure responsible, ethical, and human-centered recommendation practices. The key contributions of this research address the following: balancing accuracy and diversity and promoting fairness while ensuring user satisfaction and engagement.

In summary, this research provides innovative approaches to developing and updating responsible RSs that can tackle the inherent challenges of traditional systems, thereby contributing to a more balanced, trustworthy, and enriching recommendation experience for users.

6.2 Research Contributions

This thesis makes three significant contributions by proposing novel models that tackle key limitations in traditional and purely ML-based RSs:

Chapter 3 introduces the DOR model, which brings several key innovations to the field of RSs and addresses the central research question of this thesis: *"How can the accuracy of RSs be improved while addressing the limitations of traditional RSs, such as*

the cold start problem and data sparsity, and maintaining responsible practices?" A key contribution of the DOR model is its dual-observation mechanism, which combines local and global observation modules to enhance both recommendation accuracy and content diversity. The local observation module captures low-order and high-order relationships within textual content, allowing for a more nuanced understanding of user preferences. The global observation module complements this by analyzing interactions between users' belief networks and external information sources. This dual approach enables the DOR model to deliver highly personalized recommendations that reflect users' evolving interests while also promoting diverse content exposure by incorporating a broader context from external sources.

A central limitation in traditional and ML-based RSs is data sparsity and the cold start problem, which can impair accurate recommendations for new users or with limited interaction data. The DOR model tackles these challenges by leveraging local and global information enriched by KGs. This combination of low-order and high-order relationship modeling allows the system to uncover complex semantic relationships that traditional models might overlook. As a result, even users with sparse historical data benefit from accurate and relevant recommendations.

Another notable contribution is the introduction of the AGGC model. This model enhances the recommendation process by integrating attention mechanisms into graph-based learning. By using graph convolution and attention layers, the AGGC model improves the extraction of high-order relationships in the user-item interaction space. This allows the DOR model to account for broader contextual information in its recommendations, leading to more precise and diverse suggestions. The AGGC model effectively balances local and global information, providing a more comprehensive understanding of user preferences.

The DOR model's effectiveness was validated through extensive experiments on three real-world datasets: MIND-small, MIND-large, and IMDB. Across all datasets, the DOR model outperformed several state-of-the-art baseline models in key metrics such as AUC, MRR, NDCG@5, and NDCG@10. These results demonstrate the model's superior ability to provide accurate and personalized recommendations. The experimental results underline the practical utility of incorporating dual observations and attention-based graph learning into recommendation tasks.

In addition to the main experimental evaluations, this chapter also presents a series of ablation studies to assess the contribution of the dual-observation mechanism, the AGGC model, and various KG embedding methods (such as TransE, TransH, and TransR). The ablation results confirm the importance of these components in enhancing the DOR model's overall performance. Further analysis was conducted to understand the impact of different

hyperparameters, such as epoch size, embedding dimensions, and filter sizes. This analysis helped optimize the model's configuration and confirmed the robustness of the proposed approach.

In conclusion, the DOR model marks a substantial step forward in overcoming the challenges faced by traditional RSs, especially in addressing data sparsity and cold start issues. Through its dual-observation mechanism and advanced graph-based techniques, the DOR model delivers more accurate, diverse, and personalized recommendations. Chapter 3 also highlights future research opportunities, including the development of more refined interaction mechanisms, enhanced representations of entity relationships, and the integration of additional features such as images and categories to further improve the model's real-world effectiveness.

Chapter 4 makes significant contributions toward addressing the issue of "filter bubbles" in RSs, proposing a novel RGRec framework that enhances content diversity and promotes belief harmony among users. The research result presented in this chapter addresses Research Question 2: *"How can RSs enhance diversity and mitigate filter bubbles to improve user satisfaction and trust?"* The RGRec framework combines algorithmic and human-focused strategies, distinguishing it from traditional methods that focus on either algorithmic diversity or user-centric nudging techniques. By integrating these approaches, RGRec provides a more balanced and effective solution to the persistent "filter bubble" problem.

A key contribution of this chapter is the introduction of the Multi-faceted Reasoning-based FBDetect. This module identifies users affected by "filter bubbles" by constructing belief networks, where a user's belief system is represented as a heterogeneous graph. The FBDetect module operates on two fronts: system-level bias detection and user belief bias identification, providing a comprehensive analysis of filter bubbles and enabling targeted interventions. This dual-layered analysis offers a novel approach to filter bubble detection, advancing current methodologies by offering both algorithmic and user-focused insights.

Another important contribution is the development of the Belief Nudging module, which aims to diversify users' content exposure and broaden their interests. This module identifies paths between topics that users strongly prefer and those they tend to avoid, using a graph-based similarity measure. Through this nudging process, users are subtly encouraged to engage with content outside of their usual preferences, thus moderating extreme beliefs and reducing the risk of ideological isolation. The incorporation of libertarian paternalism principles into this nudging process ensures that users' autonomy is respected, while simultaneously guiding them toward more balanced content exposure.

Furthermore, the RecomGen module is another novel contribution of this chapter. RecomGen uses large language models to generate contextually rich items based on the recom-

mendation prompt paths explored by the Belief Nudging module. By integrating generative AI with traditional RSs, RGRec provides users with personalized, diverse content recommendations that enhance their exposure to a broader range of topics, while also promoting belief harmony. The use of generative AI in this context represents a step forward in how RSs can leverage AI to combat filter bubbles more effectively.

The experimental results provide strong empirical support for the efficacy of RGRec. Through evaluations on real-world datasets, such as MIND and IMDB, the results demonstrate that RGRec significantly outperforms baseline models in terms of recommendation diversity, belief harmony, and the reduction of filter bubble-affected users. The framework proves to be adaptable, effectively enhancing user belief diversity and improving RSs' overall quality. In addition, RGRec's iterative feedback mechanism allows it to continually adjust recommendations based on user responses, making it more responsive and human-centered than many existing systems.

In summary, this research contributes to the advancement of responsible RSs by proposing an innovative framework that integrates both graph-based algorithms and human-centered nudging techniques. This approach not only addresses the technical challenge of "filter bubbles" but also contributes to the broader discourse on responsible AI by ensuring that RSs promote diversity and mitigate bias in a manner that respects user autonomy. Through strict experimentation and validation, the research establishes RGRec as a promising solution for future RSs, with implications for both academic research and practical applications in diverse domains.

Chapter 5 answers the research question *"How can RSs be designed to achieve responsible and balanced information delivery without altering existing algorithms?"* It introduces the AAIN model, a novel framework designed to mitigate "filter bubble" effects in RSs without modifying the underlying algorithms. Inspired by the traditional Chinese Yin-Yang theory, the AAIN model balances information perception by providing users with both positive and negative viewpoints, promoting a more neutral understanding of diverse topics. The core innovation lies in the introduction of a multi-agent framework, where the OPA collaborates with an AINA to provide a balanced set of recommendations. This approach allows the AAIN model to intervene selectively, neutralizing bias in the recommended content while maintaining user engagement.

One of the significant contributions of this chapter is the development of the YYNC mechanism, which plays a central role in the AAIN model by ensuring that contrasting perspectives are incorporated into the recommendations. YYNC identifies and adjusts imbalanced recommendations by clustering the content based on sentiment and semantic similarity. Through this mechanism, the model neutralizes recommendations that could

otherwise contribute to the formation of filter bubbles, thus promoting a more diverse and balanced information exposure.

Moreover, this chapter introduces an adaptive framework that allows AAIN to operate on top of existing RSs. By doing so, it addresses a critical challenge in the field: mitigating "filter bubbles" without requiring substantial changes to existing systems. This adaptive feature is particularly important because it ensures the model's practical applicability across various recommendation platforms without disrupting the core algorithms that drive user engagement.

Extensive experiments demonstrate the effectiveness of the AAIN model in enhancing recommendation diversity and achieving sentiment neutrality. The results show that the AAIN model can significantly reduce the impact of filter bubbles by broadening users' exposure to diverse viewpoints, all while maintaining high levels of recommendation accuracy. By comparing the performance of the AAIN-enhanced model with several baseline models, the chapter underscores the value of incorporating the AAIN framework into preference-based RSs. The experiments highlight the importance of cluster size in optimizing performance, revealing how different configurations affect sentiment neutralization.

In summary, this chapter contributes to the advancement of responsible RSs by proposing a model that effectively balances content diversity without compromising user preferences. It demonstrates the viability of a multi-agent, Yin-Yang-inspired framework for mitigating "filter bubbles" and provides a robust experimental validation of the model's performance across multiple datasets. The AAIN model thus represents a significant step forward in designing RSs that are both user-centric and socially responsible.

These models collectively represent significant contributions toward developing responsible RSs. By addressing specific challenges related to data sparsity, "filter bubbles", and bias, they offer practical, scalable solutions for improving both recommendation quality and user experience.

In summary, this thesis presents substantial contributions to the development of responsible RSs by addressing major challenges such as accuracy, diversity, and fairness. Through the introduction of the DOR, RGRec, and AAIN models, it tackles critical issues like data sparsity, cold start problems, filter bubbles, and imbalanced information delivery. These models collectively demonstrate the potential for RSs to not only enhance personalized content delivery but also promote a more balanced and fair information environment. Together, they pave the way for future research and practical applications, ensuring that RSs evolve in a manner that is both user-centered and socially responsible.

6.3 Limitations and Future Directions

While this thesis introduces significant advancements in the development of RSs, certain limitations remain that warrant further investigation. The proposed models have demonstrated effectiveness in controlled environments, but their robustness and scalability need more thorough testing in diverse real-world settings. Such testing will help to better understand how these models perform in varying user contexts and dynamic environments.

Moreover, the integration of user-driven customization options within the models could provide richer insights into individual preferences for information diversity. Allowing users to adjust the level of diversity and neutrality in the recommendations might lead to greater satisfaction and engagement. Future research will focus on addressing the following areas:

1. **Scalability in Real-World Settings:** The proposed models have been validated in controlled environments, but their scalability and performance in real-world, dynamic contexts need further investigation.

2. **User-Centered Customization:** Future work could explore more personalized user-driven customization options, allowing users to control the level of diversity and neutrality in their recommendations, potentially leading to higher satisfaction and engagement.

3. **Long-Term Impact on User Behavior:** It is important to investigate the long-term effects of these models on users' beliefs and behaviors to understand their broader social implications.

4. **Integration of Multimedia Content:** Incorporating multimedia elements, such as images and videos, into RSs could further diversify content and cater to a wider range of user preferences, enhancing the overall recommendation experience.

These areas represent promising directions for future research, which could extend the practical application and social impact of responsible RSs.

6.4 Summary

In conclusion, this thesis presents novel contributions to developing RSs that address long-standing challenges, including cold start, data sparsity, filter bubbles, and fairness. The three proposed models, DOR, RGRec, and AAIN, provide innovative solutions that enhance

accuracy, diversity, and fairness, ensuring users receive balanced and trustworthy recommendations. These models form a strong foundation for creating ethical, RSs capable of adapting to evolving user needs while promoting a fair and inclusive information ecosystem.

The research presented in this thesis demonstrates the potential for responsible RSs to deliver high-quality, personalized recommendations and address the ethical challenges posed by bias and information diversity. These contributions open new pathways for future research and practical applications in building more user-centered recommendation technologies.

References

[1] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: a system for large-scale machine learning. In *Osdi*, volume 16, pages 265–283. Savannah, GA, USA.

[2] Afoudi, Y., Lazaar, M., and Al Achhab, M. (2021). Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network. *Simulation Modelling Practice and Theory*, 113:102375.

[3] Aggarwal, C. C. et al. (2018). Neural networks and deep learning. *Springer*, 10:978–3.

[4] Al Jawarneh, I. M., Bellavista, P., Corradi, A., Foschini, L., Montanari, R., Berrocal, J., and Murillo, J. M. (2020). A pre-filtering approach for incorporating contextual information into deep learning based recommender systems. *IEEE Access*, 8:40485–40498.

[5] Al-Mani, I. A., Al-Sabaawi, A. M. A., and Hussien, M. H. (2022). A review paper of model based collaborative filtering techniques. In *2022 International Conference on Data Science and Intelligent Computing (ICDSIC)*, pages 52–57. IEEE.

[6] Alatawi, F., Cheng, L., Tahir, A., Karami, M., Jiang, B., Black, T., and Liu, H. (2021). A survey on echo chambers on social media: Description, detection and mitigation.

[7] Alhijawi, B., Obeid, N., Awajan, A., and Tedmori, S. (2022). New hybrid semantic-based collaborative filtering recommender systems. *International Journal of Information Technology*, 14(7):3449–3455.

[8] Amara, S. and Subramanian, R. R. (2020). Collaborating personalized recommender system and content-based recommender system using textcorpus. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 105–109. IEEE.

[9] Amrollahi, A. et al. (2021). A conceptual tool to eliminate filter bubbles in social networks. *Australasian journal of information systems*, 25.

[10] An, M., Wu, F., Wu, C., Zhang, K., Liu, Z., and Xie, X. (2019). Neural news recommendation with long-and short-term user representations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 336–345.

[11] Anderson, A., Maystre, L., Anderson, I., Mehrotra, R., and Lalmas, M. (2020). Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of the web conference 2020*, pages 2155–2165.

[12] Anwar, T. and Uma, V. (2019). Mrec-crm: Movie recommendation based on collaborative filtering and rule mining approach. In *2019 international conference on Smart Structures and Systems (ICSSS)*, pages 1–5. IEEE.

[13] Areeb, Q. M., Nadeem, M., Sohail, S. S., Imam, R., Doctor, F., Himeur, Y., Hussain, A., and Amira, A. (2023). Filter bubbles in recommender systems: Fact or fallacy—a systematic review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1512.

[14] Aridor, G., Goncalves, D., and Sikdar, S. (2020). Deconstructing the filter bubble: User decision-making and recommender systems. In *Proceedings of the 14th ACM conference on recommender systems*, pages 82–91.

[15] Babu, M. S. P. and Kumar, B. R. S. (2011). An implementation of the user-based collaborative filtering algorithm. *IJCSIT) International Journal of Computer Science and Information Technologies*, 2(3):1283–1286.

[16] Bansal, M., Goyal, A., and Choudhary, A. (2022). A comparative analysis of k-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal*, 3:100071.

[17] Barbehenn, M. (1998). A note on the complexity of dijkstra's algorithm for graphs with weighted vertices. *IEEE transactions on computers*, 47(2):263.

[18] Bechmann, A. and Nielbo, K. L. (2018). Are we exposed to the same "news" in the news feed? *Digital Journalism*, 6(8):990–1002.

[19] Behera, D. K., Das, M., Swetanisha, S., and Sethy, P. K. (2021). Hybrid model for movie recommendation system using content k-nearest neighbors and restricted boltzmann machine. *Indonesian Journal of Electrical Engineering and Computer Science*, 23(1):445–452.

[20] Behrendt, M. and Harmeling, S. (2021). Arguebert: How to improve bert embeddings for measuring the similarity of arguments. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 28–36.

[21] Belavadi, P., Burbach, L., Halbach, P., Nakayama, J., Plettenberg, N., Ziefle, M., and Valdez, A. C. (2020). Filter bubbles and content diversity? an agent-based modeling approach. In *International Conference on Human-Computer Interaction*, pages 215–226. Springer.

[22] Bellina, A., Castellano, C., Pineau, P., Iannelli, G., and De Marzo, G. (2023). Effect of collaborative-filtering-based recommendation algorithms on opinion polarization. *Physical Review E*, 108(5):054304.

[23] Benabbes, K., Housni, K., El Mezouary, A., and Zellou, A. (2022). Recommendation system issues, approaches and challenges based on user reviews. *Journal of Web Engineering*, 21(4):1017–1054.

[24] Beshears, J. and Kosowsky, H. (2020). Nudging: Progress to date and future directions. *Organizational behavior and human decision processes*, 161:3–19.

[25] Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., et al. (2019). Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2212–2220.

[26] Bhanuse, R. and Mal, S. (2021). A systematic review: deep learning based e-learning recommendation system. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 190–197. IEEE.

[27] Blumenthal-Barby, J. and Opel, D. J. (2018). Nudge or grudge? choice architecture and parental decision-making. *Hastings Center Report*, 48(2):33–39.

[28] Brundha, J. and Meera, K. (2022). Vector model based information retrieval system with word embedding transformation. In *2022 10th International Conference on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP-22)*, pages 01–04. IEEE.

[29] Bryant, L. V. (2020a). The youtube algorithm and the alt-right filter bubble. *Open Information Science*, 4(1):85–90.

[30] Bryant, L. V. (2020b). The youtube algorithm and the alt-right filter bubble. *Open Information Science*, 4(1):85–90.

[31] Calvillo, D. P., Swan, A. B., and Rutchick, A. M. (2020). Ideological belief bias with political syllogisms. *Thinking & Reasoning*, 26(2):291–310.

[32] Cao, X., Shi, Y., Yu, H., Wang, J., Wang, X., Yan, Z., and Chen, Z. (2021). Dekr: description enhanced knowledge graph for machine learning method recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 203–212.

[33] Cao, Y., Wang, X., He, X., Hu, Z., and Chua, T.-S. (2019). Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences. In *The world wide web conference*, pages 151–161.

[34] Castells, P., Hurley, N., and Vargas, S. (2021). Novelty and diversity in recommender systems. In *Recommender systems handbook*, pages 603–646. Springer.

[35] Cen, Y., Zhang, J., Zou, X., Zhou, C., Yang, H., and Tang, J. (2020). Controllable multi-interest framework for recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2942–2951.

[36] Chaitanya, V. S., Mohan, M., and Thilagam, P. S. (2022). A clustering-based model for the generation of diversified recommendations. In *2022 IEEE Silchar Subsection Conference (SILCON)*, pages 1–6. IEEE.

[37] Channarong, C., Paosirikul, C., Maneeroj, S., and Takasu, A. (2022). Hybridbert4rec: a hybrid (content-based filtering and collaborative filtering) recommender system based on bert. *IEEE Access*, 10:56193–56206.

[38] Chen, C., Zhang, M., Zhang, Y., Liu, Y., and Ma, S. (2020a). Efficient neural matrix factorization without sampling for recommendation. *ACM Transactions on Information Systems (TOIS)*, 38(2):1–28.

[39] Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., and He, X. (2023). Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3):1–39.

[40] Chen, K. and Ma, R. (2016). A mathematic expression of the genes of chinese traditional philosophy. *arXiv e-prints*, pages arXiv–1611.

[41] Chen, M., Ma, T., and Zhou, X. (2022). Cocnn: Co-occurrence cnn for recommendation. *Expert Systems with Applications*, 195:116595.

[42] Chen, X., Jia, S., and Xiang, Y. (2020b). A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141:112948.

[43] Chicaiza, J. and Valdiviezo-Diaz, P. (2021). A comprehensive survey of knowledge graph-based recommender systems: Technologies, development, and contributions. *Information*, 12(6):232.

[44] Chitra, U. and Musco, C. (2020). Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 115–123.

[45] Chueca Del Cerro, C. (2024). The power of social networks and social media's filter bubble in shaping polarisation: an agent-based model. *Applied Network Science*, 9(1):69.

[46] Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., and Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the national academy of sciences*, 118(9):e2023301118.

[47] Cinus, F., Minici, M., Monti, C., and Bonchi, F. (2022). The effect of people recommenders on echo chambers and polarization. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 90–101.

[48] Daelemans, W. and Van Den Bosch, A. (2010). Memory-based learning. *The Handbook of Computational Linguistics and Natural Language Processing*, pages 154–179.

[49] Dahlgren, P. M. (2021). A critical review of filter bubbles and a comparison with selective exposure. *Nordicom Review*, 42(1):15–33.

[50] Dalecke, S. and Karlsen, R. (2020). Designing dynamic and personalized nudges. In *Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics*, pages 139–148.

[51] Daniels, Z. A., Frank, L. D., Menart, C. J., Raymer, M., and Hitzler, P. (2020). A framework for explainable deep neural models using external knowledge graphs. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II*, volume 11413, pages 480–499. SPIE.

[52] Defazio, A. and Jelassi, S. (2022). Adaptivity without compromise: a momentumized, adaptive, dual averaged gradient method for stochastic optimization. *J Mach Learn Res*, 23:1–34.

[53] Del Olmo, F. H. and Gaudioso, E. (2008). Evaluation of recommender systems: A new approach. *Expert Systems with Applications*, 35(3):790–804.

[54] Deldjoo, Y., Anelli, V. W., Zamani, H., Bellogin, A., and Di Noia, T. (2021). A flexible framework for evaluating user and item fairness in recommender systems. *User Modeling and User-Adapted Interaction*, pages 1–55.

[55] Dennehy, D., Griva, A., Pouloudi, N., Dwivedi, Y. K., Mäntymäki, M., and Pappas, I. O. (2023). Artificial intelligence (ai) and information systems: perspectives to responsible ai. *Information Systems Frontiers*, 25(1):1–7.

[56] Deshpande, M. and Karypis, G. (2004). Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 22(1):143–177.

[57] Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., de Prado, M. L., Herrera-Viedma, E., and Herrera, F. (2023). Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation. *Information Fusion*, 99:101896.

[58] Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*, volume 1. Springer.

[59] Do, H.-Q., Le, T.-H., and Yoon, B. (2020). Dynamic weighted hybrid recommender systems. In *2020 22nd International Conference on Advanced Communication Technology (ICACT)*, pages 644–650. IEEE.

[60] Dongliang, Z., Yi, W., and Zichen, W. (2022). Review of recommendation systems based on knowledge graph. *Data analysis and knowledge discovery*, 5(12):1–13.

[61] Donkers, T. and Ziegler, J. (2021). The dual echo chamber: Modeling social media polarization for interventional recommending. In *Proceedings of the 15th ACM conference on recommender systems*, pages 12–22.

[62] Dooms, S., De Pessemier, T., and Martens, L. (2011). A user-centric evaluation of recommender algorithms for an event recommendation system. In *RecSys 2011 Workshop on Human Decision Making in Recommender Systems (Decisions@ RecSys' 11) and User-Centric Evaluation of Recommender Systems and Their Interfaces-2 (UCERSTI 2) affiliated with the 5th ACM Conference on Recommender Systems (RecSys 2011)*, pages 67–73. Ghent University, Department of Information technology.

[63] Duan, H., Liu, P., and Ding, Q. (2023). Rfan: Relation-fused multi-head attention network for knowledge graph enhanced recommendation. *Applied Intelligence*, 53(1):1068–1083.

[64] Duong, T. N., Vuong, T. A., Nguyen, D. M., and Dang, Q. H. (2020). Utilizing an autoencoder-generated item representation in hybrid recommendation system. *IEEE Access*, 8:75094–75104.

[65] Ekström, A. G., Niehorster, D. C., and Olsson, E. J. (2022). Self-imposed filter bubbles: Selective attention and exposure in online search. *Computers in Human Behavior Reports*, 7:100226.

[66] Elahi, E., Anwar, S., Shah, B., Halim, Z., Ullah, A., Rida, I., and Waqas, M. (2024). Knowledge graph enhanced contextualized attention-based network for responsible user-specific recommendation. *ACM Transactions on Intelligent Systems and Technology*.

[67] Elahi, M., Jannach, D., Skjærven, L., Knudsen, E., Sjøvaag, H., Tolonen, K., Holmstad, Ø., Pipkin, I., Throndsen, E., Stenbom, A., et al. (2022). Towards responsible media recommendation. *AI and Ethics*, pages 1–12.

[68] Ertz, M., Jo, M.-S., Karakas, F., and Sarigöllü, E. (2021). Message sidedness effects in advertising: the role of yin-yang balancing theory. *Social Sciences*, 10(6):229.

[69] Fan, H., Zhong, Y., Zeng, G., and Ge, C. (2022). Improving recommender system via knowledge graph based exploring user preference. *Applied Intelligence*, pages 1–13.

[70] Fan, W., Ma, Y., Li, Q., He, Y., Zhao, E., Tang, J., and Yin, D. (2019). Graph neural networks for social recommendation. In *The world wide web conference*, pages 417–426.

[71] Fensel, D., Şimşek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, I., Umbrich, J., Wahler, A., Fensel, D., et al. (2020). Introduction: what is a knowledge graph? *Knowledge graphs: Methodology, tools and selected use cases*, pages 1–10.

[72] Flaxman, S., Goel, S., and Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320.

[73] Gao, Z., Shen, T., Mai, Z., Bouadjenek, M. R., Waller, I., Anderson, A., Bodkin, R., and Sanner, S. (2022). Mitigating the filter bubble while maintaining relevance: Targeted diversification with vae-based recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2524–2531.

[74] Gazdar, A. and Hidri, L. (2020). A new similarity measure for collaborative filtering based recommender systems. *Knowledge-Based Systems*, 188:105058.

[75] Ge, Y., Zhao, S., Zhou, H., Pei, C., Sun, F., Ou, W., and Zhang, Y. (2020). Understanding echo chambers in e-commerce recommender systems. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 2261–2270.

[76] Geetha, G., Safa, M., Fancy, C., and Saranya, D. (2018). A hybrid approach using collaborative filtering and content based filtering for recommender system. In *Journal of physics: conference series*, volume 1000, page 012101. IOP Publishing.

[77] Gharahighehi, A. and Vens, C. (2021). Personalizing diversity versus accuracy in session-based recommender systems. *SN Computer Science*, 2(1):39.

[78] Gharahighehi, A. and Vens, C. (2023). Diversification in session-based news recommender systems. *Personal and Ubiquitous Computing*, 27(1):5–15.

[79] Ghazanfar, M. A. and Prugel-Bennett, A. (2010). A scalable, accurate hybrid recommender system. In *2010 Third International Conference on Knowledge Discovery and Data Mining*, pages 94–98. IEEE.

[80] Giannopoulos, G., Weber, I., Jaimes, A., and Sellis, T. (2012). Diversifying user comments on news articles. In Wang, X. S., Cruz, I., Delis, A., and Huang, G., editors, *Web Information Systems Engineering - WISE 2012*, pages 100–113, Berlin, Heidelberg. Springer Berlin Heidelberg.

[81] Gohari, F. and Tarokh, M. (2017). Classification and comparison of the hybrid collaborative filtering systems. *International journal of research in industrial engineering*, 6(2):129–148.

[82] Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., and Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100.

[83] Grossetti, Q., Du Mouza, C., and Travers, N. (2019). Community-based recommendations on twitter: avoiding the filter bubble. In *Web Information Systems Engineering–WISE 2019*, pages 212–227.

[84] Guo, Q., Sun, Z., Zhang, J., and Theng, Y.-L. (2020a). An attentional recurrent neural network for personalized next location recommendation. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 34, pages 83–90.

[85] Guo, Q., Zhuang, F., Qin, C., Zhu, H., Xie, X., Xiong, H., and He, Q. (2020b). A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3549–3568.

[86] Guo, Z., Yu, K., Li, Y., Srivastava, G., and Lin, J. C.-W. (2021). Deep learning-embedded social internet of things for ambiguity-aware social recommendations. *IEEE Transactions on Network Science and Engineering*.

[87] Gupta, S. and Dave, M. (2020). An overview of recommendation system: methods and techniques. *Advances in Computing and Intelligent Systems: Proceedings of ICACM 2019*, pages 231–237.

[88] Haider, M. (2024). Political polarization in the digital age: Understanding social dynamics. *Physical Education, Health and Social Sciences*, 2(1):19–30.

[89] Hao, Y., Dong, L., Wei, F., and Xu, K. (2019). Visualizing and understanding the effectiveness of BERT. 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)*, pages 4143–4152.

[90] Hayashi, K. (2022). Rethinking correlation-based item-item similarities for recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2287–2291.

[91] He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., and Wang, M. (2020). Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648.

[92] Hu, Y., Wu, S., Jiang, C., Li, W., Bai, Q., and Roehrer, E. (2022a). Ai facilitated isolations? the impact of recommendation-based influence diffusion in human society. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 5080–5086.

[93] Hu, Y., Wu, S., Jiang, C., Li, W., Bai, Q., and Roehrer, E. (2022b). Ai facilitated isolations? the impact of recommendation-based influence diffusion in human society. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, pages 5080–5086.

[94] Hui, B., Zhang, L., Zhou, X., Wen, X., and Nian, Y. (2022). Personalized recommendation system based on knowledge embedding and historical behavior. *Applied Intelligence*, pages 1–13.

[95] Isinkaye, F. O., Folajimi, Y. O., and Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal*, 16(3):261–273.

[96] Jain, G., Mahara, T., and Tripathi, K. N. (2020). A survey of similarity measures for collaborative filtering-based recommender system. In *Soft Computing: Theories and Applications: Proceedings of SoCTA 2018*, pages 343–352. Springer.

[97] Javed, U., Shaukat, K., Hameed, I. A., Iqbal, F., Alam, T. M., and Luo, S. (2021). A review of content-based and context-based recommendation systems. *International Journal of Emerging Technologies in Learning (iJET)*, 16(3):274–306.

[98] Jeong, C., Jang, S., Park, E., and Choi, S. (2020). A context-aware citation recommendation model with bert and graph convolutional networks. *Scientometrics*, 124:1907–1922.

[99] Jesse, M. and Jannach, D. (2021). Digital nudging with recommender systems: Survey and future directions. *Computers in Human Behavior Reports*, 3:100052.

[100] Jiang, N., Gao, L., Duan, F., Wen, J., Wan, T., and Chen, H. (2022). San: Attention-based social aggregation neural networks for recommendation system. *International Journal of Intelligent Systems*, 37(6):3373–3393.

[101] Jiang, X. (2013). Chinese dialectical thinking—the yin yang model. *Philosophy Compass*, 8(5):438–446.

[102] Jin, R. and Zhou, L. (2021). An improved knowledge representation model based on transh. In *2021 2nd International Conference on Artificial Intelligence and Information Systems*, pages 1–6.

[103] Joachim, S., Forkan, A. R. M., Jayaraman, P. P., Morshed, A., and Wickramasinghe, N. (2022). A nudge-inspired ai-driven health platform for self-management of diabetes. *Sensors*, 22(12):4620.

[104] Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM.

[105] Jung, J., Son, B., and Lyu, S. (2020). Attnio: Knowledge graph exploration with in-and-out attention flow for knowledge-grounded dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3484–3497.

[106] Kandel, A. and Zhang, Y.-Q. (1998). Intrinsic mechanisms and application principles of general fuzzy logic through yin-yang analysis. *Information Sciences*, 106(1-2):87–104.

[107] Karetnikov, A., Ehrlinger, L., and Geist, V. (2022). Enhancing transe to predict process behavior in temporal knowledge graphs. In *Database and Expert Systems Applications-DEXA 2022 Workshops: 33rd International Conference, DEXA 2022, Vienna, Austria, August 22–24, 2022, Proceedings*, pages 369–374. Springer.

[108] Kenton, J. D. M.-W. C. and Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, page 2.

[109] Khanal, S. S., Prasad, P., Alsadoon, A., and Maag, A. (2020). A systematic review: machine learning based recommendation systems for e-learning. *Education and Information Technologies*, 25(4):2635–2664.

[110] Kitchens, B., Johnson, S. L., and Gray, P. (2020). Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption. *MIS quarterly*, 44(4).

[111] Ko, H., Lee, S., Park, Y., and Choi, A. (2022). A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141.

[112] Krichene, W. and Rendle, S. (2022). On sampled metrics for item recommendation. *Communications of the ACM*, 65(7):75–83.

[113] Kumar, B., Sharma, N., and Sharma, S. (2020). Collaborative topic regression-based recommendation systems: A comparative study. In *Proceedings of ICRIC 2019: Recent Innovations in Computing*, pages 723–737. Springer.

[114] Kumar, P. S. (2020). Recommendation system for e-commerce by memory based and model based collaborative filtering. In *Proceedings of the 11th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2019)*, volume 1182, page 123. Springer.

[115] Kuznetsov, S. and Kordík, P. (2023). Overcoming the cold-start problem in recommendation systems with ontologies and knowledge graphs. In *European Conference on Advances in Databases and Information Systems*, pages 591–603. Springer.

[116] Lei, F., Liu, X., Dai, Q., Ling, B. W.-K., Zhao, H., and Liu, Y. (2020). Hybrid low-order and higher-order graph convolutional networks. *Computational Intelligence and Neuroscience*, 2020.

[117] Li, D., Qu, H., and Wang, J. (2023a). A survey on knowledge graph-based recommender systems. In *2023 China Automation Congress (CAC)*, pages 2925–2930. IEEE.

[118] Li, G., Zhu, J., and Xi, H. (2021a). Deep recommendation based on dual attention mechanism. In *2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 675–680. IEEE.

[119] Li, J., Zhu, J., Bi, Q., Cai, G., Shang, L., Dong, Z., Jiang, X., and Liu, Q. (2022). Miner: Multi-interest matching network for news recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 343–352.

[120] Li, L., Zhang, Z., and Zhang, S. (2021b). Hybrid algorithm based on content and collaborative filtering in recommendation system optimization and simulation. *Scientific Programming*, 2021(1):7427409.

[121] Li, W. (2018). *Comprehensive Modelling of Influence Diffusion in Complex Social Networks, an Agent-based Perspective*. PhD thesis, Auckland University of Technology.

[122] Li, W., Bai, Q., Nguyen, T. D., and Zhang, M. (2017). Agent-based influence maintenance in social networks. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1592–1594.

[123] Li, W., Bai, Q., and Zhang, M. (2019a). Siminer: A stigmergy-based model for mining influential nodes in dynamic social networks. *IEEE Transactions on Big Data*, 5(2):223–237.

[124] Li, W., Bai, Q., Zhang, M., and Nguyen, T. D. (2019b). Automated Influence Maintenance in Social Networks: an Agent-based Approach. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1884–1897.

[125] Li, Y., Chen, H., Fu, Z., Ge, Y., and Zhang, Y. (2021c). User-oriented fairness in recommendation. In *Proceedings of the web conference 2021*, pages 624–632.

[126] Li, Z., Dong, Y., Gao, C., Zhao, Y., Li, D., Hao, J., Zhang, K., Li, Y., and Wang, Z. (2023b). Breaking filter bubble: A reinforcement learning framework of controllable recommender system. In *Proceedings of the ACM Web Conference 2023*, pages 4041–4049.

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

[128] Liu, G. and An, R. (2021). Applying a yin–yang perspective to the theory of paradox: a review of chinese management. *Psychology research and behavior management*, pages 1591–1601.

[129] Liu, R., Yin, G., Liu, Z., and Zhang, L. (2023a). Ptke: Translation-based temporal knowledge graph embedding in polar coordinate system. *Neurocomputing*, 529:80–91.

[130] Liu, Z., Xu, W., Zhang, W., and Jiang, Q. (2023b). An emotion-based personalized music recommendation framework for emotion improvement. *Information Processing & Management*, 60(3):103256.

[131] Lops, P., Polignano, M., Musto, C., Silletti, A., and Semeraro, G. (2023). Clayrs: An end-to-end framework for reproducible knowledge-aware recommender systems. *Information Systems*, 119:102273.

[132] Lunardi, G. M., Machado, G. M., Maran, V., and de Oliveira, J. P. M. (2020a). A metric for filter bubble measurement in recommender algorithms considering the news domain. *Applied Soft Computing*, 97:106771.

[133] Lunardi, G. M., Machado, G. M., Maran, V., and de Oliveira, J. P. M. (2020b). A metric for filter bubble measurement in recommender algorithms considering the news domain. *Applied Soft Computing*, 97.

[134] Ma, M., Na, S., Wang, H., Chen, C., and Xu, J. (2022). The graph-based behavior-aware recommendation for interactive news. *Applied Intelligence*, 52(2):1913–1929.

[135] Mak, H., Koprinska, I., and Poon, J. (2003). Intimate: A web-based movie recommender using text categorization. In *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, pages 602–605. IEEE.

[136] Martins, G. B., Papa, J. P., and Adeli, H. (2020). Deep learning techniques for recommender systems based on collaborative filtering. *Expert Systems*, 37(6):e12647.

[137] Michiels, L., Leysen, J., Smets, A., and Goethals, B. (2022). What are filter bubbles really? a review of the conceptual and empirical work. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 274–279.

[138] Million, E. (2007). The hadamard product. *Course Notes*, 3(6):1–7.

[139] Mohamed, S. K., Nounu, A., and Nováček, V. (2021). Biological applications of knowledge graph embedding models. *Briefings in bioinformatics*, 22(2):1679–1693.

[140] Mooney, R. J. and Roy, L. (2000). Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 195–204.

[141] Mourthé, A. and Mello, C. E. (2022). Less is more: improving neural-based collaborative filtering by using landmark modeling. *Information Sciences*, 590:217–233.

[142] Munawar, S., Ali, Z., Waqas, M., Tu, S., Hassan, S. A., and Abbas, G. (2022). Cooperative computational offloading in mobile edge computing for vehicles: A model-based dnn approach. *IEEE Transactions on Vehicular Technology*, 72(3):3376–3391.

[143] Musto, C., Basile, P., and Semeraro, G. (2019). Hybrid semantics-aware recommendations exploiting knowledge graph embeddings. In *AI* IA 2019–Advances in Artificial Intelligence: XVIIIth International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, November 19–22, 2019, Proceedings 18*, pages 87–100. Springer.

[144] Nguyen, L. V., Hong, M.-S., Jung, J. J., and Sohn, B.-S. (2020a). Cognitive similarity-based collaborative filtering recommendation system. *Applied Sciences*, 10(12):4183.

[145] Nguyen, L. V., Nguyen, T.-H., and Jung, J. J. (2020b). Content-based collaborative filtering using word embedding: a case study on movie recommendation. In *Proceedings of the international conference on research in adaptive and convergent systems*, pages 96–100.

[146] Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., and Konstan, J. A. (2014). Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pages 677–686.

[147] Niu, Z., Zhong, G., and Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62.

[148] Noorian, A., Harounabadi, A., and Ravanmehr, R. (2022). A novel sequence-aware personalized recommendation system based on multidimensional information. *Expert Systems with Applications*, 202:117079.

[149] Noulapeu Ngaffo, A. and Choukair, Z. (2022). A deep neural network-based collaborative filtering using a matrix factorization with a twofold regularization. *Neural computing and applications*, 34(9):6991–7003.

[150] Omran, P. G., Wang, K., and Wang, Z. (2019). Learning temporal rules from knowledge graph streams. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*.

[151] Pandit, R. and Naskar, S. K. (2015). A memory based approach to word sense disambiguation in bengali using k-nn method. In *2015 IEEE 2nd international conference on recent trends in information systems (ReTIS)*, pages 383–386. IEEE.

[152] Pariser, E. (2011). *The filter bubble: what the internet is hiding from you*. Penguin UK, London, UK.

[153] Patro, G. K., Biswas, A., Ganguly, N., Gummadi, K. P., and Chakraborty, A. (2020). Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *Proceedings of the web conference 2020*, pages 1194–1204.

[154] Pedronette, D. C. G. and Torres, R. d. S. (2012). Exploiting pairwise recommendation and clustering strategies for image re-ranking. *Information Sciences*, 207:19–34.

[155] Peng, K. and Nisbett, R. E. (1999). Culture, dialectics, and reasoning about contradiction. *American psychologist*, 54(9):741.

[156] Pisner, D. A. and Schnyer, D. M. (2020). Support vector machine. In *Machine learning*, pages 101–121. Elsevier.

[157] Qiu, Z., Wu, X., Gao, J., and Fan, W. (2021). U-bert: Pre-training user representations for improved recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4320–4327.

[158] Raghuwanshi, S. K. and Pateriya, R. (2019). Collaborative filtering techniques in recommendation systems. *Data, Engineering and Applications: Volume 1*, pages 11–21.

[159] Ramalingam, G. and Reps, T. (1993). A categorized bibliography on incremental computation. In *Proceedings of the 20th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 502–510.

[160] Ramzan, B., Bajwa, I. S., Jamil, N., Amin, R. U., Ramzan, S., Mirza, F., and Sarwar, N. (2019). An intelligent data analysis for recommendation systems using machine learning. *Scientific Programming*, 2019(1):5941096.

[161] Rawat, A., Ghildiyal, S., Dixit, A. K., Memoria, M., Kumar, R., and Kumar, S. (2022). Approaches towards ai-based recommender system. In *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, volume 1, pages 191–196. IEEE.

[162] Reddy, S., Nalluri, S., Kunisetti, S., Ashok, S., and Venkatesh, B. (2019). Content-based movie recommendation system using genre correlation. In *Smart Intelligent Computing and Applications: Proceedings of the Second International Conference on SCI 2018, Volume 2*, pages 391–397.

[163] Ren, W.-Q., Qu, Y.-B., Dong, C., Jing, Y.-Q., Sun, H., Wu, Q.-H., and Guo, S. (2023). A survey on collaborative dnn inference for edge intelligence. *Machine Intelligence Research*, 20(3):370–395.

[164] Resnick, P., Garrett, R. K., Kriplean, T., Munson, S. A., and Stroud, N. J. (2013). Bursting your (filter) bubble: strategies for promoting diverse exposure. In *Proceedings of the 2013 conference on Computer supported cooperative work companion*, pages 95–100.

[165] Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., and Meira Jr, W. (2020). Auditing radicalization pathways on youtube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 131–141.

[166] Ricci, F., Rokach, L., and Shapira, B. (2021). Recommender systems: Techniques, applications, and challenges. *Recommender systems handbook*, pages 1–35.

[167] Rostami, M., Oussalah, M., and Farrahi, V. (2022). A novel time-aware food recommender-system based on deep learning and graph clustering. *IEEE Access*, 10:52508–52524.

[168] Sagdic, A., Tekinbas, C., Arslan, E., and Kucukyilmaz, T. (2020). A scalable k-nearest neighbor algorithm for recommendation system problems. In *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 186–191. IEEE.

[169] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295.

[170] Schumann, C., Foster, J., Mattei, N., and Dickerson, J. (2020). We need fairness and explainability in algorithmic hiring. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

[171] Seridi, K. and El Rharras, A. (2023). A comparative analysis of memory-based and model-based collaborative filtering on recommender system implementation. In *The Proceedings of the International Conference on Smart City Applications*, pages 75–86. Springer.

[172] Shah, L., Gaudani, H., and Balani, P. (2016). Survey on recommendation system. *International Journal of Computer Applications*, 137(7):43–49.

[173] Sharma, S., Rana, V., and Kumar, V. (2021). Deep learning based semantic personalized recommendation system. *International Journal of Information Management Data Insights*, 1(2):100028.

[174] Sharma, S., Rana, V., and Malhotra, M. (2022). Automatic recommendation system based on hybrid filtering algorithm. *Education and Information Technologies*, pages 1–16.

[175] Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *The Journal of Experimental Education*, 61(4):293–316.

[176] Shekhar, S., Singh, A., and Gupta, A. K. (2022). A deep neural network (dnn) approach for recommendation systems. In *Advances in Computational Intelligence and Communication Technology: Proceedings of CICT 2021*, pages 385–396. Springer.

[177] Sheu, H.-S. and Li, S. (2020). Context-aware graph embedding for session-based news recommendation. In *Fourteenth ACM conference on recommender systems*, pages 657–662.

[178] Sheugh, L. and Alizadeh, S. H. (2015). A note on pearson correlation coefficient as a metric of similarity in recommender system. In *2015 AI & Robotics (IRANOPEN)*, pages 1–6. IEEE.

[179] Shi, J., Li, W., Yongchareon, S., Yang, Y., and Bai, Q. (2022). Graph-based joint pandemic concern and relation extraction on twitter. *Expert Systems with Applications*, 195:116538.

[180] Shlezinger, N., Whang, J., Eldar, Y. C., and Dimakis, A. G. (2023). Model-based deep learning. *Proceedings of the IEEE*, 111(5):465–499.

[181] Shokrzadeh, Z., Feizi-Derakhshi, M.-R., Balafar, M.-A., and Mohasefi, J. B. (2024). Knowledge graph-based recommendation system enhanced by neural collaborative filtering and knowledge graph embedding. *Ain Shams Engineering Journal*, 15(1):102263.

[182] Silveira, T., Zhang, M., Lin, X., Liu, Y., and Ma, S. (2019). How good your recommender system is? a survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, 10:813–831.

[183] Sinha, B. B. and Dhanalakshmi, R. (2022). Dnn-mf: Deep neural network matrix factorization approach for filtering information in multi-criteria recommender systems. *Neural Computing and Applications*, 34(13):10807–10821.

[184] Sitar-Tăut, D.-A., Mican, D., and Buchmann, R. A. (2021). A knowledge-driven digital nudging approach to recommender systems built on a modified onicescu method. *Expert Systems with Applications*, 181:115170.

[185] Sodan, A. C. (1998). Yin and yang in computer science. *Communications of the ACM*, 41(4):103–114.

[186] Son, J. and Kim, S. B. (2017). Content-based filtering for recommendation systems using multiattribute networks. *Expert Systems with Applications*, 89:404–412.

[187] Sonboli, N., Smith, J. J., Cabral Berenfus, F., Burke, R., and Fiesler, C. (2021). Fairness and transparency in recommendation: The users' perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 274–279.

[188] Srba, I., Moro, R., Tomlein, M., Pecher, B., Simko, J., Stefancova, E., Kompan, M., Hrckova, A., Podrouzek, J., Gavornik, A., et al. (2023). Auditing youtube's recommendation algorithm for misinformation filter bubbles. *ACM Transactions on Recommender Systems*, 1(1):1–33.

[189] Stegmann, D., Magin, M., and Stark, B. (2022). Filter bubbles. *Elgar Encyclopedia of Technology and Politics*, 37(2):220.

[190] Steinskog, D. J., Tjøstheim, D. B., and Kvamstø, N. G. (2007). A cautionary note on the use of the kolmogorov–smirnov test for normality. *Monthly Weather Review*, 135(3):1151–1157.

[191] Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., and Jiang, P. (2019). Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450.

[192] Sun, Z. and Luo, N. (2010). A new user-based collaborative filtering algorithm combining data-distribution. In *2010 International Conference of Information Science and Management Engineering*, volume 2, pages 19–23. IEEE.

[193] Sun, Z., Yang, J., Zhang, J., Bozzon, A., Huang, L.-K., and Xu, C. (2018). Recurrent knowledge graph embedding for effective recommendation. In *Proceedings of the 12th ACM conference on recommender systems*, pages 297–305.

[194] Sundararajan, L. (2014). The function of negative emotions in the confucian tradition. *The positive side of negative emotions*, pages 179–197.

[195] Tahmasebi, F., Meghdadi, M., Ahmadian, S., and Valiallahi, K. (2021). A hybrid recommendation system based on profile expansion technique to alleviate cold start problem. *Multimedia Tools and Applications*, 80:2339–2354.

[196] Tang, H., Wu, S., Xu, G., and Li, Q. (2023a). Dynamic graph evolution learning for recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1589–1598.

[197] Tang, J., Shen, S., Wang, Z., Gong, Z., Zhang, J., and Chen, X. (2023b). When fairness meets bias: a debiased framework for fairness aware top-n recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 200–210.

[198] Trigeorgis, G., Bousmalis, K., Zafeiriou, S., and Schuller, B. W. (2016). A deep matrix factorization method for learning attribute representations. *IEEE transactions on pattern analysis and machine intelligence*, 39(3):417–429.

[199] Verma, P. and Sharma, S. (2020). Artificial intelligence based recommendation system. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 669–673. IEEE.

[200] Vilela, A. L., Pereira, L. F. C., Dias, L., Stanley, H. E., and Silva, L. R. d. (2021). Majority-vote model with limited visibility: an investigation into filter bubbles. *Physica A: Statistical Mechanics and its Applications*, 563.

[201] Wang, D., Wang, X., Xiang, Z., Yu, D., Deng, S., and Xu, G. (2021a). Attentive sequential model based on graph neural network for next poi recommendation. *World Wide Web*, 24(6):2161–2184.

[202] Wang, D., Xu, D., Yu, D., and Xu, G. (2021b). Time-aware sequence model for next-item recommendation. *Applied Intelligence*, 51:906–920.

[203] Wang, D., Yih, Y., and Ventresca, M. (2020a). Improving neighbor-based collaborative filtering by using a hybrid similarity measurement. *Expert Systems with Applications*, 160:113651.

[204] Wang, F., Zhu, H., Srivastava, G., Li, S., Khosravi, M. R., and Qi, L. (2021c). Robust collaborative filtering recommendation with user-item-trust records. *IEEE Transactions on Computational Social Systems*, 9(4):986–996.

[205] Wang, G., Li, W., Bai, Q., and Lai, E. M.-K. (2023a). Maximizing social influence with minimum information alteration. *IEEE Transactions on Emerging Topics in Computing*, pages 1–13.

[206] Wang, H., Amagata, D., Makeawa, T., Hara, T., Hao, N., Yonekawa, K., and Kurokawa, M. (2020b). A dnn-based cross-domain recommender system for alleviating cold-start problem in e-commerce. *IEEE Open Journal of the Industrial Electronics Society*, 1:194–206.

[207] Wang, H., Hong, Z., and Hong, M. (2022a). Research on product recommendation based on matrix factorization models fusing user reviews. *Applied Soft Computing*, 123:108971.

[208] Wang, H., Wu, F., Liu, Z., and Xie, X. (2020c). Fine-grained interest matching for neural news recommendation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 836–845.

[209] Wang, H., Zhang, F., Wang, J., Zhao, M., Li, W., Xie, X., and Guo, M. (2018a). Ripplenet: Propagating user preferences on the knowledge graph for recommender systems. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 417–426.

[210] Wang, H., Zhang, F., Xie, X., and Guo, M. (2018b). Dkn: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 world wide web conference*, pages 1835–1844.

[211] Wang, H., Zhang, F., Zhao, M., Li, W., Xie, X., and Guo, M. (2019a). Multi-task feature learning for knowledge graph enhanced recommendation. In *The world wide web conference*, pages 2000–2010.

[212] Wang, H., Zhao, M., Xie, X., Li, W., and Guo, M. (2019b). Knowledge graph convolutional networks for recommender systems. In *The world wide web conference*, pages 3307–3313.

[213] Wang, M., Hu, Y., Wu, S., Li, W., Bai, Q., Yuan, Z., and Jiang, C. (2024). Nudging towards responsible recommendations: a graph-based approach to mitigate belief filter bubbles. *IEEE Transactions on Artificial Intelligence*, pages 1–15.

[214] Wang, M., Li, W., Shi, J., Wu, S., and Bai, Q. (2023b). Dor: a novel dual-observation-based approach for recommendation systems. *Applied Intelligence*, 53(23):29109–29127.

[215] Wang, M., Ren, P., Mei, L., Chen, Z., Ma, J., and De Rijke, M. (2019c). A collaborative session-based recommendation approach with parallel memory modules. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 345–354.

[216] Wang, R., Wu, Z., Lou, J., and Jiang, Y. (2022b). Attention-based dynamic user modeling and deep collaborative filtering recommendation. *Expert Systems with Applications*, 188:116036.

[217] Wang, S., Hu, L., Wang, Y., Sheng, Q. Z., Orgun, M., and Cao, L. (2019d). Modeling multi-purpose sessions for next-item recommendations via mixture-channel purpose routing networks. In *International joint conference on artificial intelligence*. International Joint Conferences on Artificial Intelligence.

[218] Wang, W., Feng, F., Nie, L., and Chua, T.-S. (2022c). User-controllable recommendation against filter bubbles. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1251–1261.

[219] Wang, W., Xu, Y., Feng, F., Lin, X., He, X., and Chua, T.-S. (2023c). Diffusion recommender model. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 832–841.

[220] Wang, X., He, X., Cao, Y., Liu, M., and Chua, T.-S. (2019e). Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 950–958.

[221] Wang, X., He, X., Wang, M., Feng, F., and Chua, T.-S. (2019f). Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, pages 165–174.

[222] Wang, X., Jin, H., Zhang, A., He, X., Xu, T., and Chua, T.-S. (2020d). Disentangled graph collaborative filtering. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1001–1010.

[223] Wei, W., Wang, Z., Fu, C., Damaševičius, R., Scherer, R., and Wožniak, M. (2020). Intelligent recommendation of related items based on naive bayes and collaborative filtering combination model. In *Journal of Physics: conference Series*, volume 1682, page 012043. IOP Publishing.

[224] Wong, P. T. (2013). Toward a dual-systems model of what makes life worth living. In *The human quest for meaning*, pages 3–22. Routledge.

[225] Wooditch, A., Johnson, N. J., Solymosi, R., Medina Ariza, J., and Langton, S. (2021). The normal distribution and single-sample significance tests. *A Beginner's Guide to Statistics for Criminology and Criminal Justice Using R*, pages 155–168.

[226] Wu, C., Wu, F., An, M., Huang, J., Huang, Y., and Xie, X. (2019a). Neural news recommendation with attentive multi-view learning. *arXiv preprint arXiv:1907.05576*.

[227] Wu, C., Wu, F., An, M., Huang, J., Huang, Y., and Xie, X. (2019b). Npa: neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2576–2584.

[228] Wu, C., Wu, F., Ge, S., Qi, T., Huang, Y., and Xie, X. (2019c). Neural news recommendation with multi-head self-attention. 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)*, pages 6389–6394.

[229] Wu, C., Wu, F., Huang, Y., and Xie, X. (2022). Personalized news recommendation: Methods and challenges. *ACM Transactions on Information Systems (TOIS)*.

[230] Wu, C., Wu, F., Qi, T., and Huang, Y. (2020a). Sentirec: Sentiment diversity-aware neural news recommendation. In *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing*, pages 44–53.

[231] Wu, J., Wang, X., Feng, F., He, X., Chen, L., Lian, J., and Xie, X. (2021). Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 726–735.

[232] Wu, S., Bai, Q., and Sengvong, S. (2018). Greencommute: An influence-aware persuasive recommendation approach for public-friendly commute options. *Journal of Systems Science and Systems Engineering*, 27(2):250–264.

[233] Wu, S., Li, W., and Bai, Q. (2023a). Gac: A deep reinforcement learning model toward user incentivization in unknown social networks. *Knowledge-Based Systems*, 259:110060.

[234] Wu, S., Li, W., Shen, H., and Bai, Q. (2023b). Identifying influential users in unknown social networks for adaptive incentive allocation under budget restriction. *Information Sciences*, 624:128–146.

[235] Wu, S., Xu, G., and Wang, X. (2023c). Soac: Supervised off-policy actor-critic for recommender systems. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 1421–1426.

[236] Wu, Z., Li, C., Cao, J., and Ge, Y. (2020b). On scalability of association-rule-based recommendation: A unified distributed-computing framework. *ACM Transactions on the Web (TWEB)*, 14(3):1–21.

[237] Wynne, H. E. and Wint, Z. Z. (2019). Content based fake news detection using n-gram models. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*, pages 669–673.

[238] Xiao, F. (2019). Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy. *Information Fusion*, 46:23–32.

[239] Xu, C., Zhao, P., Liu, Y., Sheng, V. S., Xu, J., Zhuang, F., Fang, J., and Zhou, X. (2019). Graph contextualized self-attention network for session-based recommendation. In *IJCAI*, volume 19, pages 3940–3946.

[240] Xu, W.-R., Lin, H.-S., Chen, X.-Y., and Zhang, Y. (2011). Yin-yang balance therapy on regulating cancer stem cells. *Journal of Traditional Chinese Medicine*, 31(2):158–160.

[241] Yadav, N., Mundotiya, R. K., Singh, A. K., and Pal, S. (2021). Diversity in recommendation system: A cluster based approach. In *Hybrid Intelligent Systems: 19th International Conference on Hybrid Intelligent Systems (HIS 2019) held in Bhopal, India, December 10-12, 2019 19*, pages 113–122. Springer.

[242] Yang, L., Wang, S., Tao, Y., Sun, J., Liu, X., Yu, P. S., and Wang, T. (2023). Dgrec: Graph neural network for recommendation with diversified embedding generation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 661–669.

[243] Yang, Y., Huang, C., Xia, L., and Li, C. (2022). Knowledge graph contrastive learning for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1434–1443.

[244] Yao, N., Liu, Q., Li, X., Yang, Y., and Bai, Q. (2022). Entity similarity-based negative sampling for knowledge graph embedding. In *PRICAI 2022: Trends in Artificial Intelligence: 19th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2022, Shanghai, China, November 10–13, 2022, Proceedings, Part II*, pages 73–87. Springer.

[245] Yu, C., Lakshmanan, L., and Amer-Yahia, S. (2009). It takes variety to make a world: diversification in recommender systems. In *Proceedings of the 12th international conference on extending database technology: Advances in database technology*, pages 368–378.

[246] Yuan, Z., Li, W., and Bai, Q. (2023). An assessment of the influence of interaction and recommendation approaches on the formation of information filter bubbles. In *Pacific Rim Knowledge Acquisition Workshop*, pages 98–110. Springer.

[247] Yue, W., Wang, Z., Tian, B., Pook, M., and Liu, X. (2020). A hybrid model-and memory-based collaborative filtering algorithm for baseline data prediction of friedreich's ataxia patients. *IEEE Transactions on Industrial Informatics*, 17(2):1428–1437.

[248] Zhang, J., Wang, L., Chen, X., Jiang, Y., and Tian, Z. (2023a). Artificial intelligence-based education platform course recommendation system. In *Proceedings of the 4th International Conference on Artificial Intelligence and Computer Engineering*, pages 835–841.

[249] Zhang, L., Li, X., Li, W., Zhou, H., and Bai, Q. (2021a). Context-aware recommendation system using graph-based behaviours analysis. *Journal of Systems Science and Systems Engineering*, 30:482–494.

[250] Zhang, Q., Li, J., Jia, Q., Wang, C., Zhu, J., Wang, Z., and He, X. (2021b). Unbert: User-news matching bert for news recommendation. In *IJCAI*, volume 21, pages 3356–3362.

[251] Zhang, Q., Lu, J., and Jin, Y. (2021c). Artificial intelligence in recommender systems. *Complex & Intelligent Systems*, 7(1):439–457.

[252] Zhang, Q., Wang, R., Yang, J., and Xue, L. (2022). Structural context-based knowledge graph embedding for link prediction. *Neurocomputing*, 470:109–120.

[253] Zhang, S., Yao, L., Sun, A., and Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1):1–38.

[254] Zhang, W.-R. (2016). Information conservational yinyang bipolar quantum-fuzzy cognitive maps-mapping business data to business intelligence. In *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 2279–2286. IEEE.

[255] Zhang, X., Wang, H., and Li, H. (2023b). Disentangled representation for diversified recommendations. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 490–498.

[256] Zheng, G., Zhang, F., Zheng, Z., Xiang, Y., Yuan, N. J., Xie, X., and Li, Z. (2018). Drn: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*, pages 167–176.

[257] Zheng, P. (2022). Multisensor feature fusion-based model for business english translation. *Scientific Programming*, 2022:1–10.

[258] Zheng, W., Zhou, Y., Liu, S., Tian, J., Yang, B., and Yin, L. (2022). A deep fusion matching network semantic reasoning model. *Applied Sciences*, 12(7):3416.

[259] Zhu, J., Dai, Q., Su, L., Ma, R., Liu, J., Cai, G., Xiao, X., and Zhang, R. (2022). Bars: towards open benchmarking for recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2912–2923.

[260] Zhu, Q., Zhou, X., Song, Z., Tan, J., and Guo, L. (2019). Dan: Deep attention neural network for news recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5973–5980.