Measuring and Optimizing Beyond-accuracy Objectives in Next Basket Recommendation

Yuanna Liu

The research was carried out at the Information Retrieval Lab at the University of Amsterdam and funded by the China Scholarship Council.

Copyright © 2026 Yuanna Liu, Amsterdam, The Netherlands
Cover by Yuanna Liu
Printed by Ridderprint, Alblasserdam

ISBN: 978-94-6568-015-6

Measuring and Optimizing Beyond-accuracy Objectives in Next Basket Recommendation

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op 19 november 2026, te 10:00 uur

door

Yuanna Liu

geboren te Shaanxi

Promotiecommissie

Promotor:	prof. dr. M. de Rijke	Universiteit van Amsterdam
Co-promotor:	dr. M. Alian Nejadi	Universiteit van Amsterdam
Overige leden:	dr. M.C. Heuss	Universiteit van Amsterdam
	prof. dr. E. Kanoulas	Universiteit van Amsterdam
	dr. A. Lucic	Universiteit van Amsterdam
	prof. dr. N. Tintarev	Maastricht University
	prof. dr. S. Verberne	Universiteit Leiden

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Acknowledgements

My PhD has been the golden age of my life. I was fortunate enough to realise it in my very first year.

First of all, I would like to thank my supervisor, Prof. Maarten de Rijke. Our weekly meetings were always the highlight of my week, and they witnessed my entire journey of becoming a researcher. I truly enjoyed discussing ideas and new discoveries with you. You always responded with professionalism, helping me think more carefully, define problems more rigorously, and stay curious. Thank you for teaching me how to lead a research project. It is my honor to be your student.

I am deeply grateful to my co-supervisor, Dr. Mohammad Aliannejadi. You taught me the fundamental way of thinking about research. When I lacked confidence, you always gave me the encouragement and reassurance I needed. You inspired me to explore new research directions and possibilities, and gently pushed me to develop stronger and more rigorous arguments. Beyond research, you have also been a kind friend, with an impressive talent for making pizza.

To my paranymph, Yubao, thank you for being such a supportive collaborator. Whenever we worked together, I always felt at ease because I knew we would solve every problem together. I also treasure the movies we watched and the wonderful meals you cooked. Gabrielle, thank you for all our weekend adventures—tree climbing, indoor surfing, boating, hiking, skating, and exploring restaurants from different countries. I have always admired your courage and kindness.

When I first arrived in the Netherlands, I was curious about everything. My officemates, Maartje and Maurits, helped me in countless ways. You introduced me to ABBA, and we often shared songs and sang together after lunch in the office. These are wonderful memories for me. I would also like to thank Ming. I always enjoyed our discussions about the next basket recommendation project. I learned the importance of being rigorous and making forward-looking judgments in those meetings.

Throughout my PhD, I was surrounded by colleagues who generously offered their help. Gabriel, thank you for telling me not to let experimental results determine my mood, but to treat them like the weather in the Netherlands during winter. Thank you also for reading the book I recommended. Jingfen, Jin, and Xingying, thank you for helping me get back efficiently from Morocco. Maria, thank you for helping me improve my soos talk slides and teaching me how to prepare a good presentation. Songgaojun, thank you for your calm presence and our conversations. Mariya, Jiayi, Yixian and Jia-Hong, thank you for generously sharing your experience and advice on interview preparation. Lu, I like our unplanned weekend trips to explore nearby cities. I am also grateful to Shiguang for the ping-pong games, which brought me some happy moments. Thanks to Jingwei for hosting so many hotpot dinners and board game nights at your place. I also enjoyed organizing the fairness tutorials with Clara, Marleen and Chen. I would like to thank the students from the RecSys course; our collaborations have been both rewarding and inspiring. To all my coauthors, thank you for the discussions, the support, and the joy of working together. You have made research a journey of exploration, where even the difficulties are merely scenery along the way.

I would also like to thank my officemates, Masoud, Yongkang, Clemencia, Jing-

wei, Oscar, Panagiotis, Simon, Yixing, and Xinyi. Those small conversations, timely feedback, and everyday interactions made the office a wonderful place to work. We shared so much, both inside and outside research. I am grateful to the members of the fairness reading group: Ali, Fatemeh, Jasmin, Ana, Maria, and Clara. And the generative retrieval reading group: Kidist, Rui, Shiguang, Dominykas, and Yubao. I always enjoyed discussing papers with you.

Thanks to all the IRLab members for showing up to every soos talk and research meeting. Your presence made my PhD experience richer, more complete, and more joyful: Andrew, Antonios, Chuan, David, Dylan, Harrie, Hongyi, Ivana, Jiaxuan, Kiki, Maarten, Maik, Maxime, Mingwei, Mozhdeh, Olivier, Pablo, Petra, Philipp, Pooya, Romain, Roxana, Ruben, Ruqing, Samarth, Sami, Shaojie, Shashank, Siddharth, Teng, Thilina, Thong, Vaishali, Vera, Weijia, Yibin, Yougang, Yuyue, Zahra, and Zihan.

I would like to express my sincere gratitude to Maria, Evangelos, Ana, Nava, and Suzan for being my Ph.D. committee and for generously dedicating their time and expertise to reviewing my work.

I am grateful to my piano teacher, Francesco, for teaching me every week over the past three years. Choosing a new piece together and working through it became one of the greatest joys of my life. Thank you for your patience, which helped me find my most faithful friend, the piano. It listened to all the emotions I could not put into words, letting my joy, pain, anger, and love find an outlet. Again and again, I reminded myself not to forget those moments when a melody touched something deep within me, when I could hear its brightness and its sigh. To this day, Ryuichi Sakamoto's music remains my favorite.

A few years ago, Prof. Deng introduced me to research. Your guidance and encouragement gave me the confidence to take my first step. For the first time, I wished one day I could do the same work as you.

Last winter, the Netherlands experienced a rare heavy snowfall. The whole world turned white, with snowflakes flying through the air. The snowfall left such a deep impression on me that I have often found myself thinking back to it ever since.

My heartfelt thanks to Mama Najia. We met by pure chance, yet you welcomed me with such trust and affection that I immediately felt at home. You showed me the desert and the sea around Laayoune, cooked the most delicious food for me. Those days will remain among the rarest and most precious memories of my life.

To my old friends Tingqian and Cong, my sister Yifan, and my brother Yi, thank you for the conversations and video calls whenever I felt lonely. You always made the distance feel a little smaller and brought back the warmth and familiarity of home.

To my parents and grandparents, thank you for always believing in me, for trusting me to make my own choices, and for giving me the freedom to pursue the life I wanted. It gave me the confidence to explore the world on my own.

Finally, thank you, Ping Chen. Your writing has been part of my life since my teenage years; your courage and free spirit have shaped me in more ways than you know.

Yuanna
Amsterdam
July 2026

Contents

1 Introduction **1**
- 1.1 Research Outline and Questions . . . 3
- 1.2 Main Contributions . . . 5
  - 1.2.1 Algorithmic contributions . . . 5
  - 1.2.2 Theoretical contributions . . . 6
  - 1.2.3 Empirical contributions . . . 6
- 1.3 Thesis Overview . . . 7
- 1.4 Origins . . . 7

2 Measuring Item Fairness in Next Basket Recommendation **11**
- 2.1 Introduction . . . 11
- 2.2 Related Work . . . 13
  - 2.2.1 Next basket recommendation . . . 13
  - 2.2.2 Item fairness in recommender systems . . . 14
- 2.3 Reproducibility Setup . . . 14
  - 2.3.1 Problem formulation . . . 14
  - 2.3.2 Datasets . . . 15
  - 2.3.3 NBR methods . . . 16
  - 2.3.4 Fair ranking metrics . . . 16
  - 2.3.5 Position weighting models . . . 18
  - 2.3.6 Implementation details . . . 18
- 2.4 Experiments & Results . . . 19
- 2.5 Conclusion . . . 23
- 2.6 Appendix . . . 24
  - 2.6.1 Reproducibility . . . 24
  - 2.6.2 Additional plots . . . 24

3 Are We Really Achieving Better Beyond-Accuracy Performance in Next Basket Recommendation? **29**
- 3.1 Introduction . . . 29
- 3.2 Related Work . . . 33
- 3.3 Task Formulation and Definitions . . . 34
- 3.4 Evaluation Metrics . . . 34
- 3.5 A Two-Step Repetition-Exploration Framework . . . 36
  - 3.5.1 Repetition module . . . 37
  - 3.5.2 Exploration module . . . 38
  - 3.5.3 Basket generation module . . . 39
- 3.6 Experiments . . . 39
  - 3.6.1 Experimental setup . . . 39
  - 3.6.2 Overall accuracy performance . . . 41
  - 3.6.3 Beyond-accuracy performance . . . 43
  - 3.6.4 Reflections and discussions . . . 46
- 3.7 Conclusion . . . 47
- 3.8 Appendix . . . 48

4 Repeat-bias-aware Optimization of Beyond-Accuracy Metrics for Next Basket Recommendation **49**
4.1 Introduction . . . 49
4.2 Related Work . . . 51
4.3 Method . . . 52
4.3.1 Problem formulation . . . 52
4.3.2 Model overview . . . 53
4.3.3 Optimization objectives . . . 53
4.3.4 Repeat-bias-aware optimization for unified NBR methods . . . 54
4.3.5 Repeat-bias-aware optimization for combined NBR methods . . . 55
4.4 Experimental Setup . . . 56
4.5 Experimental Results . . . 57
4.6 Conclusion . . . 60
4.7 Appendix . . . 62
4.7.1 Reproducibility . . . 62
4.7.2 Additional results . . . 62

5 Mask-Prompted Generative Retrieval for Diverse and Exploratory Next Basket Recommendation **67**
5.1 Introduction . . . 67
5.2 Task Definition . . . 69
5.3 MPGBR . . . 70
5.3.1 Diversity-aware item tokenization . . . 70
5.3.2 Mask-prompted generative basket recommendation . . . 73
5.4 Experimental Setup . . . 74
5.5 Results . . . 76
5.6 Related Work . . . 85
5.6.1 Next basket recommendation . . . 85
5.6.2 Generative recommendation with semantic IDs . . . 86
5.7 Conclusion . . . 86
5.8 Appendix . . . 87

6 Conclusions **89**
6.1 Summary of Findings . . . 89
6.2 Future Work . . . 91
6.2.1 Discriminative vs. semantic-ID-based generative paradigms . . . 91
6.2.2 Fairness in recommendation . . . 92

Bibliography **95**

Summary **103**

Samenvatting **105**

1 Introduction

Recommender systems connect users with information, products, or services of potential interest, helping them explore the world and enhancing the convenience of their daily lives [41, 105]. Recommender systems have been widely deployed across numerous online platforms, including music streaming, news media and e-commerce [80, 84, 94]. By using a range of user interactions (e.g., clicks and purchases), recommendation algorithms infer users' preferences and accordingly recommend relevant items to them.

Recent advances in machine learning and deep learning have been widely employed in various recommendation tasks, primarily to improve accuracy [26, 42, 53, 70, 106]. However, recommender systems are multi-objective systems that connect and serve multiple stakeholders, users, providers, and platforms, determining both how much exposure each provider receives and what information reaches users [93]. A comprehensive evaluation that goes beyond accuracy can therefore help us assess the true performance of recommenders and their potential social impact, revealing what current metrics actually reward and what they may obscure [13, 82].

Diversity is a beyond-accuracy objective that has been studied for many years. Briefly, it concerns how dissimilar the items within a list are from each other. Early work in search built on the maximal marginal relevance (MMR) reranking criterion [11]. In recommendation, it often focuses on topic diversification [154], aiming to reduce redundancy and improve the user experience. Subsequently, researchers recognized that the observational data underlying recommendation are systematically biased, and that naively trained models inherit and even amplify these biases [15, 81]. A prominent example is popularity bias, the tendency of algorithms to disproportionately favor popular items and reinforce this skew through feedback loops [23, 92]. Fairness has emerged as one of the key concerns in trustworthy recommender systems [41]. It assesses whether certain groups of users are treated discriminatorily [68], or whether certain items receive unfair exposure allocation [31, 104]. A central difficulty in item fairness is that fairness itself lacks a unified definition [121], and different definitions can lead to contradictory conclusions [100]. The two most common notions, statistical parity and equal opportunity, differ fundamentally in whether they condition on relevance (merit): the former distributes exposure equally across item groups regardless of relevance, whereas the latter allocates exposure in proportion to relevance [9, 98, 104]. Consequently, rather than seeking a universally correct fairness metric, the appropriate choice depends on the specific task and the reliability of the relevance labels. Crucially, these beyond-accuracy

objectives are not independent: some overlap in their very definitions, while others are mutually entangled and influence one another [52]. Together, they help build sustainable recommender systems while mitigating the potential harms that algorithms may pose to society. This is especially important in an era in which people increasingly delegate information gathering and decision-making to algorithmic agents [38].

To balance accuracy with beyond-accuracy objectives, numerous methods have been proposed at different stages of the recommendation pipeline [95, 127]. In-processing methods embed these objectives directly into training by augmenting the learning objective with dedicated constraints or regularization terms [72, 133, 153], attaining stronger trade-offs at the cost of retraining the recommendation model. Post-processing methods re-rank items from a given set of relevance scores [11, 16, 86], making them model-agnostic and easy to deploy, though their beyond-accuracy gains are inherently bounded by the quality of the candidate set. As the recommendation paradigm shifts, these optimization methods should be adapted accordingly. The recently emerging semantic-ID-based generative paradigm, for instance, first constructs semantic IDs for items from their metadata, then trains a generative model to autoregressively generate the target item's semantic ID conditioned on the user's historical interactions [99, 120]. This generative formulation, however, reshapes how bias and unfairness emerge, and both these mechanisms and the optimization methods designed to mitigate them remain at an early stage of exploration [34, 73].

Beyond-accuracy objectives apply to general recommendation scenarios, but they must be adapted to specific tasks to account for their particular characteristics. In retail, for instance, user behavior exhibits both repetitive and exploratory patterns. This pattern is particularly important in next basket recommendation (NBR), which predicts the next set of items given a user's historical basket sequences [45]. Numerous studies have employed frequency-based methods, k-nearest neighbor (KNN)-based methods [33, 45] and deep-learning-based methods [57, 137, 138] to improve the overall accuracy of basket recommendation. These methods target three core modeling challenges: learning a basket representation [44, 137], capturing intra-basket item correlations [57, 66, 138], and modeling inter-basket sequential dynamics [44, 139]. Li et al. [63] evaluate the accuracy of repeat items and explore items separately, along with their respective proportions in the recommended baskets. This work has given rise to an interesting finding: current state-of-the-art methods tend to exhibit a strong repeat bias, i.e., they recommend an excessive number of repeat items, with the majority of their accuracy contributed by these repeat items. Since repeat recommendation relies primarily on modeling item frequency and recency, whereas explore recommendation must infer novel items from a much larger item catalog by exploiting item-to-item correlations, recent work has begun to decouple NBR into repetitive [3, 54] and exploratory [61] NBR separately.

However, the evaluation and optimization of beyond-accuracy objectives in NBR remain largely underexplored. In particular, how the joint evaluation of accuracy and beyond-accuracy performance is shaped by the interplay between repetitive and exploratory recommendations remains an open problem. In this thesis, we study the *evaluation and optimization* of beyond-accuracy objectives, including item fairness, diversity, and repeat bias, in the context of the NBR task, which poses unique challenges arising from users' repetitive and exploratory purchase behaviors. We begin with

discriminative NBR methods and then extend to the generative paradigm, keeping these beyond-accuracy objectives in focus regardless of how the underlying paradigm evolves.

1.1 Research Outline and Questions

This thesis focuses on the next basket recommendation task and develops a series of contributions on the evaluation and optimization of beyond-accuracy objectives, including item fairness, diversity, and repeat bias. We scope our research by addressing four main research questions across four chapters.

We start by measuring the item fairness of next basket recommendation methods. Item fairness in recommender systems is concerned with assessing whether individual items or item groups obtain an equitable portion of exposure. The literature has shown that the sensitivity of fairness metrics to various configurations varies across different information access systems [98]. Therefore, item fairness metrics need to be systematically examined in NBR, where users' purchase behavior is more repeat-oriented than in other recommendation tasks. We formulate our first research question as follows:

RQ1 *What is the item fairness performance of current next basket recommendation methods?*

To answer RQ1, Chapter 2 first implements seven fair ranking metrics spanning the equal opportunity and statistical parity taxonomy, and then evaluates the item fairness of eight representative NBR methods under a unified experimental setting. We observe that different metrics yield inconsistent fairness rankings across these methods, but the majority of them concur that repeat-biased methods come out as fairer than explore-biased ones. Furthermore, we test the sensitivity of these item fairness metrics with respect to basket size and position weighting models. Our findings lead us to advocate expected exposure loss (EEL) and expected exposure disparity (EED) for NBR, as they prove to be highly robust to parameter configurations, while also being able to assess fairness across multiple groups. Finally, we investigate the item fairness of recommendations for user groups with varying degrees of repeat purchase behavior.

Building on the observation that, in NBR, most of the accuracy gain comes from recommending repeat items, together with our insights into item fairness, we design a method that optimizes accuracy, together with item fairness or diversity by leveraging the characteristics of repetitive and exploratory recommendation. To this end, we raise our second research question as follows:

RQ2 *How can we design a basket recommendation method that jointly targets accuracy and beyond-accuracy objectives, and does the current evaluation paradigm adequately support this goal?*

To answer RQ2, Chapter 3 proposes a simple yet effective strategy: repeat items can be exploited to secure strong accuracy, while explore items can be used to improve item fairness and diversity. To test this hypothesis, we propose a two-step repetition-exploration (TREx) framework that decouples repeat and explore recommendation. The repetition module captures a user's personal purchase frequency together with item-specific repurchase features, whereas the exploration module targets beyond-accuracy

objectives, namely item fairness or diversity. Then, we combine these two lists to form the final basket. Our experiments show that TREx does achieve better accuracy and beyond-accuracy performance in NBR. However, we do not simply declare success and stop there; instead, we critically examine the accuracy of explore items, which turns out to be quite low. Therefore, we diagnose a loophole in the current overall evaluation paradigm for NBR, which masks the poor accuracy of exploratory recommendation. This is unreasonable and unfair to users who prefer exploring novel items. To refine this evaluation paradigm, we argue that any joint evaluation of accuracy and beyond-accuracy metrics in NBR should be accompanied by a separate assessment of repeat items and exploratory items. In this way, we can obtain a comprehensive understanding of the true performance of NBR methods.

The above two research questions focus on the evaluation of item fairness and diversity, and on how the joint accuracy and beyond-accuracy evaluation paradigm should be adapted to account for the repeat and explore characteristics of NBR. However, another serious problem is repeat bias, i.e., NBR methods are heavily biased toward recommending repeat items in order to maximize accuracy. In our empirical study, we find that optimizing diversity or item fairness alone, without accounting for repeat bias, may lead NBR algorithms to recommend even more repeat items. We therefore introduce the following research question:

RQ3 *How can we design algorithms to jointly improve diversity and item fairness while mitigating repeat bias in next basket recommendation?*

To answer RQ3, Chapter 4 proposes two mixed-integer linear programming (MILP)-based algorithms, repeat-bias-aware diversity optimization (RADiv) and repeat-bias-aware item fairness optimization (RAIF), which re-rank the item candidate lists produced by various NBR baselines to seek a balance between repeat and explore items while optimizing for beyond-accuracy metrics. Furthermore, we extend our algorithms to another NBR paradigm [65], namely one that combines a repeat item list and an explore item list from separate models. Our experiments demonstrate that the RADiv and RAIF algorithms effectively enhance diversity and item fairness while alleviating repeat bias at an acceptable cost in utility. Repeat bias is an inescapable confounder in evaluating both accuracy and beyond-accuracy performance. We therefore argue that how to define and quantify it, particularly what proportion of repeat versus explore items is reasonable, remains an open question worth investigating.

As pointed out above, there is a growing trend to decompose NBR into repeat-only and explore-only tasks. In particular, exploratory next basket recommendation aims to surface novel and relevant items, thereby supporting item discovery and promoting long-term user engagement [17]. Meanwhile, the semantic-ID-based generative recommendation paradigm has achieved notable success in sequential recommendation: items are indexed by semantic IDs in a semantically shared manner, and a generative recommender is then trained to generate the target ID tokens conditioned on a user's history [99]. Compared with discriminative recommendation, this paradigm offers better scalability and generalization to cold-start or unseen items, since semantically similar items share overlapping ID tokens rather than being treated as atomic, independent IDs [148]. Motivated by these developments, we explore the semantic-ID-based generative paradigm for exploratory NBR, and propose the following research question:

RQ4 *How can we design a semantic-ID-based generative recommendation method for exploratory basket recommendation, and how to improve recommendation diversity under the generative paradigm?*

To answer RQ4, Chapter 5 proposes mask-prompted generative basket recommendation (MPGBR) for diverse and exploratory NBR. First, we design a mask-prompted generative framework that encodes historical basket sequences with a bidirectional transformer and recovers the semantic IDs of masked items through a lightweight decoder. We adopt a two-stage pretraining–finetuning scheme: the first stage enables the model to capture item correlations, while the second stage is tuned for future basket recommendation. In addition, we enhance recommendation diversity at the semantic IDs construction stage. Specifically, we form basket-aware semantic IDs by incorporating item co-occurrence signals into item representations, and propose Balanced RQ-KMeans, which improves recommendation diversity by lifting code assignment entropy. Experiments demonstrate that MPGBR attains accuracy comparable to that of strong baselines while lying on the accuracy–diversity Pareto frontier across all datasets. Ablation studies further reveal that Balanced RQ-KMeans is the primary driver of the diversity gain, improving diversity at the representation stage without compromising accuracy.

Overall, through an in-depth investigation of the evaluation and optimization of item fairness, diversity, and repeat bias in NBR, this thesis advances the joint evaluation of accuracy and beyond-accuracy objectives for NBR and offers valuable insights for designing beyond-accuracy-aware optimization algorithms, spanning the transition from discriminative recommendation to the generative paradigm.

1.2 Main Contributions

In this section, we list the main contributions of this thesis.

1.2.1 Algorithmic contributions

- A repetition recommendation model (TREx-Rep) that incorporates item repurchase features and users' purchase frequency (Chapter 3).
- A flexible two-step repetition–exploration framework (TREx) for NBR that controls the trade-off between accuracy and beyond-accuracy metrics over the recommended baskets (Chapter 3).
- A methodology for repeat-bias-aware re-ranking for unified NBR methods, formulating two model-agnostic MILP algorithms (RADiv and RAIF) that jointly optimize relevance, diversity (or item fairness), and repeat ratio (Chapter 4).
- An extension of RADiv and RAIF to combined NBR methods, where repeat and explore items come from separate models (Chapter 4).
- A methodology for diversity-aware item tokenization, comprising a collaborative semantic modulator and a Balanced RQ-KMeans algorithm that balances codeword assignments via optimal transport (Chapter 5).

- A mask-prompted generative framework MPGBR for exploratory basket recommendation that autoregressively generates semantic IDs from the hidden states of masked items (Chapter 5).

1.2.2 Theoretical contributions

- An analysis of the generalizability of existing fair ranking metrics under the influence of users' repetitive and exploratory purchase behaviors in next basket recommendation (Chapter 2).
- A methodology for disentangling repetition and exploration in NBR evaluation, revealing whether "better" accuracy and beyond-accuracy performance is genuine or merely a short-cut artifact (Chapter 3).
- An analysis of how optimizing diversity or item fairness alone can aggravate repeat bias in NBR, establishing the need to treat repeat bias as a joint optimization objective (Chapter 4).
- A conceptual insight that recommendation diversity in generative NBR can be optimized at the semantic ID representation stage (Chapter 5).

1.2.3 Empirical contributions

- An empirical evaluation of item exposure fairness across representative NBR methods, covering both equal opportunity and statistical parity criteria (Chapter 2).
- An empirical analysis of the sensitivity of fair ranking metrics with respect to varying basket sizes and different position weighting models (Chapter 2).
- An empirical analysis of the item fairness of recommendations across user subgroups with varying degrees of repeat purchase behavior (Chapter 2).
- An empirical accuracy comparison of TREx-Rep against baselines on two real-world grocery datasets (Chapter 3).
- An empirical analysis of the short-cut strategy via TREx across eight beyond-accuracy metrics (Chapter 3).
- An empirical comparison of beyond-accuracy metrics grouped by their connection with accuracy (strong/weak/none), showing that a stronger connection makes the short-cut harder to exploit (Chapter 3).
- An empirical analysis of exploration accuracy, revealing that TREx's apparent superiority masks low exploration accuracy and motivating fine-grained evaluation (Chapter 3).
- An empirical evaluation of RADiv and RAIF on three real-world grocery datasets, showing that both algorithms improve diversity and item fairness while mitigating repeat bias at an acceptable Recall loss (Chapter 4).

- An empirical analysis of the trade-offs between Recall and diversity, item fairness, and repeat bias (Chapter 4).
- An empirical evaluation of MPGBR on three public grocery datasets against discriminative and generative baselines, showing competitive accuracy while lying on the accuracy-diversity Pareto frontier (Chapter 5).
- An empirical analysis of balance-aware assignment across semantic ID layers, which pinpoints the deepest codebook as the primary driver of diversity gains at minimal accuracy cost (Chapter 5).
- An empirical comparison of the accuracy-diversity trade-off between Balanced RQ-KMeans and diverse beam search over their hyperparameter spaces (Chapter 5).
- An empirical analysis of how the distributional properties of semantic IDs, in particular codeword assignment entropy, drive recommendation diversity (Chapter 5).

1.3 Thesis Overview

This section provides an overview of the thesis and guidance on how to navigate it. The thesis contains six chapters, of which Chapters 2–5 investigate the research questions outlined in Section 1.1. We recommend reading the chapters in order, as they follow a coherent pipeline around the NBR task: (i) *evaluation* — Chapter 2 presents a reproducibility study of measuring item fairness in NBR, and Chapter 3 jointly evaluates accuracy and beyond-accuracy metrics (including item fairness and diversity) for NBR and identifies a loophole in the current overall-basket evaluation paradigm; (ii) *optimization in the discriminative paradigm* — Chapter 4 jointly optimizes repeat bias with beyond-accuracy objectives via re-ranking; and (iii) *optimization in the generative paradigm* — Chapter 5 proposes a generative exploratory basket recommendation method and optimizes recommendation diversity at semantic IDs construction stage. Finally, Chapter 6 summarizes the main findings of the thesis, and reflects on its limitations and promising directions for future work.

1.4 Origins

Chapter 2 is based on the following paper:

- Y. Liu, M. Li, M. Ariannezhad, M. Mansoury, M. Aliannejadi, and M. de Rijke. Measuring item fairness in next basket recommendation: A reproducibility study. In *Proceedings of the 46th European Conference on Information Retrieval*, pages 210–225, 2024.

  YL: Conceptualization, Methodology, Software, Investigation, Data Curation, Formal Analysis, Visualization, Writing – Original Draft Preparation.

ML: Resources, Formal Analysis, Writing – Review & Editing. MAr: Software, Writing – Review & Editing. MM: Methodology, Writing – Review & Editing. MAl: Supervision, Writing – Review & Editing. MdR: Supervision, Writing – Review & Editing.

Chapter 3 is based on the following paper:

- M. Li, Y. Liu, S. Jullien, M. Ariannezhad, A. Yates, M. Aliannejadi, and M. de Rijke. Are we really achieving better beyond-accuracy performance in next basket recommendation? In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 924–934, 2024.

  ML and YL contributed equally to this research. ML: Conceptualization, Methodology, Investigation, Formal Analysis, Visualization, Writing – Original Draft Preparation. YL: Conceptualization, Methodology, Software, Investigation, Formal Analysis, Visualization, Writing – Review & Editing. SJ: Writing – Review & Editing. MAr: Writing – Review & Editing. AY: Writing – Review & Editing. MAl: Supervision, Writing – Review & Editing. MdR: Supervision, Writing – Review & Editing.

Chapter 4 is based on the following paper:

- Y. Liu, M. Li, M. Aliannejadi, and M. de Rijke. Repeat-bias-aware optimization of beyond-accuracy metrics for next basket recommendation. In *Proceedings of the 47th European Conference on Information Retrieval*, pages 214–229, 2025.

  YL: Conceptualization, Methodology, Software, Investigation, Data Curation, Formal Analysis, Visualization, Writing – Original Draft Preparation. ML: Conceptualization, Writing – Review & Editing. MA: Supervision, Methodology, Writing – Review & Editing. MdR: Supervision, Writing – Review & Editing.

Chapter 5 is based on the following paper:

- Y. Liu, Y. Tang, M. Li, C. Xu, M. Aliannejadi, and M. de Rijke. Mask-prompted generative retrieval for diverse and exploratory next basket recommendation. Under review, 2026.

  YL: Conceptualization, Methodology, Software, Investigation, Data Curation, Formal Analysis, Visualization, Writing – Original Draft Preparation. YT: Methodology, Formal Analysis, Writing – Review & Editing. ML: Formal Analysis, Writing – Review & Editing. CX: Formal Analysis, Writing – Review & Editing. MA: Supervision, Writing – Review & Editing. MdR: Supervision, Writing – Review & Editing.

The writing of the thesis also benefited from work on the following publications:

- J. Hutter, H. C. Bakker, S. Fris, M. Bernardy, and Y. Liu. A systematic reproducibility study of BSARec for sequential recommendation. In *Proceedings of the 48th European Conference on Information Retrieval*, pages 17–32, 2026.

- H. Nguyen, Y. Liu, M. Mansoury, M. Aliannejadi, A. Hanjalic, and M. de Rijke. A reproducibility study of product-side fairness in bundle recommendation. In *Proceedings of the 19th ACM Conference on Recommender Systems*, pages 696–705, 2025.

- C. Rus, J. Kareem, C. Xu, Y. Liu, Z. Deng, and M. Heuss. AMS42 at the NTCIR-18 FairWeb-2 task. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, pages 61–67, 2025.

- C. Xu, Z. Deng, C. Rus, X. Ye, Y. Liu, J. Xu, Z. Dou, J.-R. Wen, and M. de Rijke. FairDiverse: A comprehensive toolkit for fairness- and diversity-aware information retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3540–3550, 2025.

- C. Xu, C. Rus, Y. Liu, M. de Jonge, J. Xu, and M. de Rijke. Fairness in information retrieval from an economic perspective. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4126–4129, 2025.

- C. Xu, C. Rus, Y. Liu, M. de Jonge, J. Xu, and M. de Rijke. Economic perspectives on fairness in information retrieval. In *Proceedings of the 48th European Conference on Information Retrieval*, pages 19–26, 2026.

- C. Xu, C. Rus, Y. Liu, M. de Jonge, J. Xu, and M. de Rijke. Towards a responsible web: Economic perspectives on fairness in information retrieval. In *Companion Proceedings of the ACM Web Conference 2026*, pages 23–26, 2026.

2

Measuring Item Fairness in Next Basket Recommendation

Item fairness of recommender systems aims to evaluate whether items receive a fair share of exposure according to different definitions of fairness. Raj and Ekstrand [98] study multiple fairness metrics under a common evaluation framework and test their sensitivity with respect to various configurations. They find that fairness metrics show varying degrees of sensitivity towards position weighting models and parameter settings under different information access systems. Although their study considers various domains and datasets, their findings do not necessarily generalize to next basket recommendation (NBR) where users exhibit a more repeat-oriented behavior compared to other recommendation domains. This chapter investigates fairness metrics in the NBR domain under a unified experimental setup, aiming to answer:

RQ1 *What is the item fairness performance of current next basket recommendation methods?*

2.1 Introduction

Fairness of information access systems is increasingly drawing attention [31]. Such systems are not only required to have high accuracy, but they should also be fair to both users and providers. Usually, items are exposed to users as ranked lists based on a *relevance* or *utility* score. The *exposure* of items influence users' browsing, clicking, and purchase behavior [98]. However, items may not receive fair exposure due to algorithmic or data biases [69]. To understand different notions and aspects of fairness, many fairness definitions and metrics have been proposed. Raj and Ekstrand [98] unify several fair ranking metrics under a common evaluation framework and compare them empirically using different information access tasks, viz. book recommendation and scholarly article retrieval and re-ranking. The authors focus on provider-side group fairness of ranked lists and design a sensitivity analysis to evaluate the robustness of fairness metrics w.r.t. position weighting models and parameter settings. Importantly,

This chapter was published as Y. Liu, M. Li, M. Ariannezhad, M. Mansoury, M. Aliannejadi, and M. de Rijke. Measuring item fairness in next basket recommendation: A reproducibility study. In *Proceedings of the 46th European Conference on Information Retrieval*, pages 210–225, 2024.

they find that *fairness metrics show different patterns of sensitivity for different search and recommendation tasks*. This means that the lessons learned in one search and recommendation scenario usually cannot be directly and completely transferred to another domain.

Item fairness and next basket recommendation. Recently, there has been increased interest in the task of next basket recommendation (NBR) [4]. NBR models users' preferences based on a sequence of historical sets of items (baskets, playlists, reading lists, ...) and then predicts, for each user, a set of items they are likely to purchase, listen to, or read next. The NBR task is important because it is relevant across a broad range of domains, from e-commerce and travel to education and entertainment. We are interested in item fairness in the context of NBR: is the exposure assigned by an NBR system fairly distributed and do the main findings from [98] generalize to the NBR task? In particular, we focus on the following findings from [98] and examine their validity in the context of NBR:

- Metrics usually disagree on the orderings of methods; we expect to find similar patterns for NBR baselines.
- For the FairTREC full retrieval task, changing the size of the ranked list has no impact on the performance of fairness measurements. But the fairness metrics attention-weighted rank fairness (AWRF) [103], IAA, EEL, EED, and expected exposure relevance (EER) are stable on the GoodReads recommendation task, while demographic parity (logDP), exposed utility ratio (logEUR), and FAIR [141] change with different-sized ranked lists and even alter the relative order of recommendation algorithms. We change the basket size and focus on the magnitude of the change of metric values and whether the fairness order of NBR methods will be affected.
- When used with different position weighting models, fairness metrics show different degrees of robustness. Position weighting models influence metric values and algorithms' ordering, especially for EEL and realized utility ratio (logRUR). For the NBR scenario, we experiment with the document-based click-through rate model (DCTR) [22], which means each position has the same exposure in a list. We expect to identify the most unstable metrics.

Item fairness and repeat behavior. Going beyond the above generalizability questions, there is a specific characteristic of NBR that may affect item fairness in unknown ways: in many NBR scenarios, users display a significant amount of repeat behavior, whereby they purchase or consume the same item multiple times. Some users mostly purchase repeat items (i.e., items that they consumed before) as part of their next basket, while others tend to explore items (items that they never bought before). Ekstrand et al. [32] suggest we need to look at metric values across different user groups rather than at the population as a whole. Following this advice, we design experiments to assess item fairness across different user groups with varying degrees of repeat purchase behavior.

To the best of our knowledge, we are the first to study fairness in next basket recommendation. We decompose RQ1 into four research questions:

(RQ1.1) What is the fairness ranking of NBR methods? Are the lessons observed in [98] valid for the NBR task?

(RQ1.2) Which fairness metrics are more robust to basket size?

(RQ1.3) Which fairness metrics are more robust to position weighting models for NBR?

(RQ1.4) How does repeat purchase behavior affect item fairness?

2.2 Related Work

2.2.1 Next basket recommendation

NBR systems make personalized product recommendations to users based on their historical baskets. Repeat purchase behavior is a prominent pattern of NBR, which makes it different from typical recommendation domains, such as movie recommendation [43]. Since the models designed for recommending repeat items and explore items vary significantly, recent work proposes more targeted task settings and algorithms [54, 60, 61] to address the unique challenges in this domain by modeling the repeat behavior of the users while trying to optimize the explore behavior of the users.

A significant amount of research concentrates on employing neural networks to learn representations for sequences of baskets. The effectiveness of recurrent neural networks (RNNs) in sequential modeling has led to their application in NBR. Yu et al. [137] propose a dynamic recurrent basket model, feeding a series of basket representations to the recurrent architecture to obtain the dynamic representation of a user. Le et al. [57] model correlation information to augment the representation of basket sequence. Yu et al. [138] propose a model based on graph neural networks (GNNs) for temporal sets prediction, where sets are constructed as weighted graphs and a graph convolutional network is applied to capture relationships among elements in each set. Ariannezhad et al. [3] analyze users' repeat consumption behavior and propose a repeat consumption-aware neural network for NBR.

However, recent nearest neighbor-based methods show more effective performance and efficiency than neural network-based baselines on NBR. Faggioli et al. [33] propose recency-aware, user-wise popularity and incorporate it into both user- and item-based collaborative approaches. Hu et al. [45] integrate temporal dynamics into personalized item frequency and then use a user-KNN method to make predictions. Naumov et al. [87] improve on [45] by considering time intervals between interactions.

Li et al. [63] reproduce NBR methods in a unified experimental setting and propose a new angle to evaluate the performance obtained from repeat items and explore items, respectively. According to whether a method tends to recommend repeat items or explore items, it is called repeat-biased or explore-biased. Our work reproduces several representative NBR methods in a unified experimental setting and focuses on measuring item fairness and exploring the impact of repeat behavior on fairness.

2.2.2 Item fairness in recommender systems

Fairness research raises challenges for information access systems characterized by (i) a multi-stakeholder nature, (ii) a rank-based problem setting, (iii) the requirement of personalization in many cases, and (iv) the role of user feedback [31]. Wang et al. [121] collect fairness definitions in the recommendation literature and provide views of classifying fairness issues. Consumer fairness cares about whether users receive comparable recommendation quality from the system [68]. In contrast, provider fairness focuses on how to assign reasonable exposure to each document, content provider, or group [98]. The allocation criteria can be with reference to a distribution [103, 131, 141] or proportional to merit [28].

Some metrics have primarily undergone testing using small and/or synthetic datasets, and they encounter challenges in handling complex real-world information access applications where incomplete data and extreme cases happen [98]. Raj and Ekstrand [98] implement experiments on book recommendation and scholarly article retrieval tasks and test the sensitivity of fairness metrics to parameter settings. Kowald et al. [55] reproduce the analyses of [1] to investigate how popularity bias causes unfairness for both long-tail items and low-mainstream users in the context of music recommendation. We can observe that the conclusions about the fairness evaluation in one domain cannot be completely transferred to another domain.

NBR has distinguishing characteristics, i.e., repeat items figure prominently amongst the recommended results and make a considerable contribution to accuracy [63]. There is no prior work studying fairness in the context of NBR. Li et al. [67] argue that the frequency bias harms the fairness of NBR system. To fill this gap, we reproduce the fairness evaluation and sensitivity experiments of [98] to (i) see if the patterns they found can be generalized to the NBR domain, and (ii) select robust metrics suitable for NBR.

2.3 Reproducibility Setup

2.3.1 Problem formulation

Our experiments concern two main parts: NBR and fairness evaluation. First, in next basket recommendation, we write D for the item set and Q for the user set. A basket is a set of items $B = [d_1, d_2, \ldots, d_m]$, where $d_i \in D$ denotes an item from the basket B. Items have no temporal order and hold equal significance in a basket. For each user $q \in Q$, there is a sequence of historical purchase baskets $B^q = [B_1^q, B_2^q, \ldots, B_n^q]$, where B_i^q indicates the i-th basket purchased by the user. The goal of NBR is to predict the next basket L for each user. Then, we evaluate the item fairness of predicted baskets among all users using the fair ranking metrics in Section 2.3.4. Specifically, the predicted basket L can be seen as a ranked list where the ranking is based on the user-specific relevance score $\hat{y}(d \mid q)$ predicted by each NBR method. The goal of item fairness is to measure whether the exposure is fairly distributed among the groups according to a specific principle. The notation used in this chapter is summarized in Table 2.1.

Table 2.1: Notation used in the chapter; adapted from [98].

Notation	Description
$d \in D$	item
$q \in Q$	user
L	ranked list of N items (predicted basket)
$L^{-1}(i)$	the item in position i of basket L
$L(d)$	rank of item d in L
$y(d\|q)$	relevance of d to q
$\hat{y}(d\|q)$	predicted relevance of d to q
$G(L)$	group alignment matrix for items in L
G^+	popular group
G^-	unpopular group
$\mathbf{a}_L$	exposure vector for items in L
ϵ_L	the exposure of groups in L $(G(L)^T \mathbf{a}_L)$

Table 2.2: Statistics of datasets after preprocessing.

Dataset	#Users	#Items	#Baskets	Avg. #baskets/user	Avg. #items/basket	Avg. repeat ratio
Instacart	19,210	29,398	305,582	15.91	10.06	0.60
Dunnhumby	2,482	37,161	107,152	43.17	10.07	0.43
TaFeng	10,182	15,023	82,387	8.09	6.14	0.21

2.3.2 Datasets

Following [3, 44, 96], we use three publicly available datasets for our experiments: (i) Instacart,[1] which contains a sample of over three million grocery orders from users. Items belonging to the same order form a basket. (ii) Dunnhumby,[2] which includes household-level transactions over two years from 2,500 households. And (iii) TaFeng,[3] which contains transaction data from a Chinese grocery store from November 2000 to February 2001. We treat all transactions of a user within a day as a basket.

For each dataset, we remove users with fewer than three baskets and items bought fewer than five times [3]. Due to the large number of baskets in the Instacart dataset, the calculation of some methods exceeds the memory, therefore we randomly sampled 20,000 users from Instacart before filtering [87]. Table 2.2 shows the statistics of three datasets after preprocessing. Avg. repeat ratio refers to the average proportion of repeat items in the ground truth baskets [63]. We split each dataset following [3, 33, 87]. The training baskets consist of all baskets of users except the last one. For users who have more than 50 baskets in the training data, we only consider their last 50 baskets in the training set [63]. The last baskets of all the users are split into 50% validation set and 50% test set.

[1] https://www.kaggle.com/datasets/yasserh/instacart-online-grocery-basket-analysis-dataset
[2] https://www.dunnhumby. com/source-files/
[3] https://www.kaggle.com/datasets/chiranjivdas09/ta-feng-grocery-dataset

Table 2.3: Summary of fairness metrics; adapted from [98].

Category	Metrics	Goal	Binomial	More Fair
Equal opportunity	logEUR	Exposure proportional to relevance	✓	○
	logRUR	Click-through rate proportional to relevance	✓	○
	IAA	Exposure proportional to predicted relevance	×	↓
	EEL,EER	Exposure matches ideal (from relevance)	×	EEL ↓, EER ↑
Statistical parity	EED	Exposure well-distributed	×	↓
	logDP	Exposure equal across groups	✓	○

2.3.3 NBR methods

We select the following representative NBR methods with open-source code, including frequency-based, nearest neighbor-based, and deep-learning-based methods. According to the classification of NBR methods in [63], G-TopFreq and Dream are explore-biased methods (the recommended baskets are skewed towards explore items averaged over all users), which are also popularity-based methods. Other methods are repeat-biased methods (the recommended baskets are skewed towards repeat items). We cover various types of NBR method and study their fairness performance.

Frequency-based methods. (i) G-TopFreq recommends the most popular k items across all historical purchases to all users. (ii) P-TopFreq counts the k products with the highest frequency in each user's historical purchase records. (iii) GP-TopFreq first recommends personal history most popular items and then fills the empty slot with global most popular items.

Nearest neighbor-based methods. (i) TIFUKNN [45] integrates temporal dynamics modeled by time-decayed weights to generate user representations. Then, the target user representation and its nearest neighbors are combined to make the prediction. (ii) UP-CF@r [33] incorporates recency-aware user-wise popularity in a collaborative filtering framework.

Deep-learning-based methods. (i) Dream [137] is an RNN-based model that learns a dynamic representation of a user and captures global sequential features among baskets. (ii) DNNTSP [138] constructs weighted graphs to learn basket-level element relationships. The attention mechanism is used to learn the temporal dependencies of sets and elements. Static and dynamic representations are fused by a gated updating mechanism. (iii) ReCANet [3] focuses on repeat consumption, combines user-item representations with historical consumption patterns, and models temporal signals by LSTM layers.

2.3.4 Fair ranking metrics

Raj and Ekstrand [98] unify several fair ranking metrics in a common framework and notation. They clarify the limitation of these metrics when applying them to practical information access systems, and test the robustness of these metrics to different design and parameter choices. We follow the notation and fairness implementation

of this paper. The fairness metrics shown in Table 2.3 are selected for the following reasons: (i) The predicted basket can be formulated as a ranking L. These metrics are well-known fairness metrics for rankings. (ii) These metrics cover two categories of fairness definitions: *statistical parity* (aimed at ensuring comparable exposure among groups) and *equal opportunity* (aimed at promoting equal treatment based on merit or utility, regardless of group membership) [98]. (iii) Since fair exposure is unlikely to be satisfied in any single ranking [9] and it is more practical to pursue fair exposure of items to overall users, we only consider the fairness of multiple rankings for NBR setting.

Assume $\pi(L \mid q)$ is a user-dependent distribution and $\rho(q)$ is a distribution over users, overall rankings among all the users follow the distribution of $\rho(q)\pi(L \mid q)$. $\epsilon_L = G(L)^T \mathbf{a}_L$ is the group exposure within a single ranking. Its expected value $\epsilon_\pi = E_{\pi\rho}[\epsilon_L]$ is the group exposure among all the rankings.

Equal opportunity. Singh and Joachims [104] propose two ratio-based metrics. Exposed utility ratio (EUR) quantifies the deviation from the objective that the exposure of each group is proportional to its utility $Y(G)$:

$$\text{EUR} = \frac{\epsilon_\pi(G^+)/Y(G^+)}{\epsilon_\pi(G^-)/Y(G^-)}. \tag{2.1}$$

Realized utility ratio (RUR) [104] models actual user engagement, the click-through rates for the groups $\Gamma(G)$ are proportional to their utility:

$$\text{RUR} = \frac{\mathbf{\Gamma}(G^+)/Y(G^+)}{\mathbf{\Gamma}(G^-)/Y(G^-)}. \tag{2.2}$$

Biega et al. [9] propose inequality of amortized attention (IAA), which takes the L_1 norm of the difference between cumulative exposure and cumulative system-predicted relevance $\hat{y}(d|q)$ for each group. For consistency, we normalize predicted relevance scores to be in the same range as exposure values:

$$\text{IAA} = \|\epsilon_\pi - \hat{Y}\|_1. \tag{2.3}$$

Diaz et al. [28] define the target exposure ϵ^* as the expected exposure under the ideal policy. Expected exposure loss (EEL) is the distance between expected exposure and target exposure:

$$\text{EEL} = \|\epsilon_\pi - \epsilon^*\|_2^2 = \|\epsilon_\pi\|_2^2 - 2{\epsilon_\pi}^T\epsilon^* + \|\epsilon^*\|_2^2. \tag{2.4}$$

EEL can be decomposed into EER $= 2{\epsilon_\pi}^T\epsilon^*$ (measuring the alignment of exposure and relevance) and expected exposure disparity (EED).

Statistical parity. EED [28] measures the inequality in exposure distribution across groups:

$$\text{EED} = \|\epsilon_\pi\|_2^2. \tag{2.5}$$

Demographic parity (DP) [104] measures the ratio of average exposure given to the two groups:

$$\text{DP} = \epsilon_\pi\left(G^+\right)/\epsilon_\pi\left(G^-\right). \tag{2.6}$$

Following [98], DP is reformulated as $\text{logDP} = \log(\epsilon_\pi(G^+) + 10^{-6}) - \log(\epsilon_\pi(G^-) + 10^{-6})$ to address the empty-group problem and enhance interpretability. logEUR and logRUR are defined in the same way.

Table 2.4: Position weighting models for computing $\mathbf{a}_L$; adapted from [98].

Metric	Model	Formula	Parameters
IAA	Geometric	$\gamma(1-\gamma)^{L(d)-1}$	patience γ
logDP, logEUR, logRUR	Logarithmic	$1/\log_2 \max\{L(d), 2\}$	–
EER, EED, EEL	RBP [83]	$\gamma^{L(d)}$	patience γ
EER, EED, EEL	Cascade	$\gamma^{L(d)-1} \prod_{j \in [0, L(d))} [1 - \phi(y(L^{-1}(j) \mid y))]$	patience γ stopping probability function ϕ
–	DCTR	$1/\lvert L(d)\rvert$	–

2.3.5 Position weighting models

Since users are likely to pay decreasing attention to lower-ranked items (position bias) [9], position weighting models are required when computing exposure [28]. These metrics explicitly represent the position weighting model as a position weight vector $\mathbf{a}_L$, as shown in Table 2.4 [98]. In NBR, many works [54, 108, 137] treat recommended results as ranked lists and evaluate NBR methods based on ranking metrics, e.g. normalized discounted cumulative gain (NDCG). Therefore, it is required to take the position of the recommended items into account when applying fair ranking metrics in NBR.

2.3.6 Implementation details

In order to evaluate group fairness using the above metrics, following previous work [40], the division of items is based on item popularity, i.e., the number of purchases in the historical baskets of all users. We define the top 20% items with the highest purchase as popular group G^+ and the remaining 80% of the items as unpopular group G^-. The default basket size is set to 10. For all NBR baselines, we perform a grid search based on the hyperparameter ranges given in the original papers to find the optimal hyperparameters using the validation set. For TIFUKNN, the number of nearest neighbors k is tuned on $\{100, 300, 500, 900, 1100, 1300\}$, the number of groups m is chosen from $\{3, 7, 11, 15, 19, 23\}$, the within-basket time-decayed ratio r_b and the group time-decayed ratio r_g are selected from $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$, and the fusion weight α is tuned on $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. For UP-CF@r, recency window r is tuned on $\{1, 5, 10, 25, 100, \infty\}$, locality q is tuned on $[1, 5, 10, 50, 100, 1000]$, and asymmetry α is tuned on $\{0, 0.25, 0.5, 0.75, 1\}$. For Dream and DNNTSP, the item embedding size is tuned on $\{16, 32, 64, 128\}$. For RecANet, user embedding size and item embedding size are tuned on $\{16, 32, 64, 128\}$. We run each method 5 times with 5 fixed random seeds to eliminate the random initialization effect and report the average results. Following [45, 63, 138], we use three accuracy metrics. Recall assesses the capacity to retrieve all relevant items. NDCG measures ranking quality, which considers sequence order by giving the lower-ranked items a discount. personalized hit ratio (PHR) computes the proportion of predicted baskets that include at least one item from the ground truth basket.

2.4 Experiments & Results

Fairness evaluation. To answer **RQ1.1**, we measure item fairness of the recommendation results obtained by each NBR method on Instacart, Dunnhumby, and TaFeng. We implement the fairness metrics following the configurations from their original papers. Table 2.5 compares the NBR methods in terms of accuracy and fairness metrics. We make the following observation: (i) In most cases, repeat-biased methods exhibit more effectiveness in Recall, NDCG, and PHR, compared to explore-biased methods. (ii) Consistent with observations in [98], our experimental results lead to different orderings of the NBR methods when ranked using the fairness metrics. For instance, on Instacart, EEL indicates DNNTSP and P-TopFreq as the top two methods, whereas logRUR ranks TopFreq and GP-TopFreq as the top two. (iii) Metrics that consider equal opportunity, namely, logEUR, logRUR, and EEL deem repeat-biased methods fairer than the explore-biased ones (G-TopFreq and Dream). Surprisingly, according to EER, G-TopFreq and Dream are fairer than the repeat-biased methods. Both G-TopFreq and Dream only recommend popular items, which is unfair. (iv) Statistical-parity metrics, namely, EED and logDP, agree that repeat-biased methods are fairer than explore-biased methods. Because statistical parity aims at ensuring equal exposures across groups, however G-TopFreq and Dream only recommend popular items.

Basket size. To answer **RQ1.2**, we change the basket size from 10 to 100 for each dataset. Figure 2.1 shows the variation of fairness values as basket size changes on Instacart.[4] It can be observed that: (i) IAA shows the highest sensitivity to basket size, as we see the highest increase in its value as the basket size grows. We relate it to the fact that as more items are considered in the list, it leads to more accumulation of gaps between position weights and predicted relevance. And the fairness order of P-TopFreq and G-TopFreq even changes. This differs from [98], where IAA exhibits stable behavior in the GoodReads recommendation task. (ii) For logDP, logEUR, and logRUR, we observe a mixed behavior for different basket sizes. As basket size increases, DNNTSP, GTop-Freq, and Dream become less fair in terms of all three metrics; PTop-Freq and ReCANet become fairer. GP-TopFreq, UF-CF@r, and TIFUKNN, however, exhibit non-monotonic changes; and these metrics reorder the NBR methods. (iii) In line with [98], we see that EEL, EED, and EER are stable with varying basket sizes.

Position weighting models. To answer **RQ1.3**, we perform experiments with different position weighting models (in Table 2.4) for each fairness metric and report the results of Instacart[5] in Figure 2.2. The parameters are assigned to the default values: patience parameter $\gamma = 0.5$, and a stopping probability of 0.5.[6] We summarize our observations below: (i) Different from [98], we observe high sensitivity of IAA on the three datasets for different position weighting models. (ii) Except for IAA, other metrics maintain the order of NBR methods across different position weighting models. Also, some slight

[4] We observe a similar trend on the Dunnhumby and TaFeng datasets and report their results in the Appendix of this chapter.

[5] Again, we observe a similar trend on the Dunnhumby and TaFeng datasets and report their results in the Appendix of this chapter.

[6] The Geometric and rank-biased precision (RBP) share the same formula under this parameter setting. Therefore, we only report the results obtained by the Geometric weighting model for fairness metrics.

Table 2.5: Overall performance comparison of NBR methods.

Data-set	Method	Accuracy			Equal opportunity					Statistical parity	
		Recall↑	NDCG↑	PHR↑	logEUR○	logRUR○	IAA↓	EEL↓	EER↑	EED↓	logDP○
Instacart	G-TopFreq	0.0704	0.0817	0.4600	4.1803	1.9140	3.8532	0.5899	3.6884	3.1169	6.9148
	Dream	0.0704	0.0817	0.4600	4.1803	$1.9143_{(0.0001)}$	$0.1466_{(0.0053)}$	0.5895	3.6879	$3.1160_{(0.0001)}$	6.9148
	P-TopFreq	0.3143	0.3339	0.8447	0.3419	0.5756	4.9492	0.0138	2.4561	1.3086	3.1112
	GP-TopFreq	0.3150	0.3343	0.8460	0.3616	0.5764	5.1071	0.0144	2.4633	1.3164	3.1313
	UP-CF@r	0.3377	0.3582	0.8586	0.7217	0.7383	5.1405	0.0351	2.4951	1.3688	3.4917
	TIFUKNN	0.3456	0.3657	0.8639	0.4275	0.5800	0.4227	0.0154	2.4095	1.2635	3.1972
	DNNTSP	$0.3295_{(0.0002)}$	$0.3434_{(0.0005)}$	$0.8581_{(0.0001)}$	$0.3458_{(0.0171)}$	$0.6050_{(0.0104)}$	$2.1511_{(0.0462)}$	$0.0020_{(0.0004)}$	$2.3760_{(0.0070)}$	$1.2166_{(0.0075)}$	$3.1154_{(0.0171)}$
	ReCANet	$0.3490_{(0.0001)}$	$0.3699_{(0.0001)}$	$0.8668_{(0.0003)}$	$0.4588_{(0.0184)}$	$0.6069_{(0.0056)}$	$2.1972_{(0.0691)}$	$0.0173_{(0.0004)}$	$2.4000_{(0.0015)}$	$1.2559_{(0.0019)}$	$3.2283_{(0.0184)}$
Dunnhumby	G-TopFreq	0.0897	0.0798	0.3795	6.2326	3.5166	2.8829	1.0020	3.1503	3.2310	6.6211
	Dream	$0.0896_{(0.0003)}$	$0.0759_{(0.0002)}$	$0.3873_{(0.0013)}$	6.2326	$3.5511_{(0.0020)}$	$0.3248_{(0.0035)}$	0.9940	3.1427	3.2154	6.6211
	P-TopFreq	0.1628	0.1562	0.5399	0.9611	1.6879	4.6402	0.4039	2.7369	2.2195	3.2579
	GP-TopFreq	0.1628	0.1562	0.5399	0.9620	1.6879	4.6480	0.4039	2.7369	2.2196	3.2588
	UP-CF@r	0.1699	0.1639	0.5536	1.2771	1.6287	4.4460	0.4279	2.7297	2.2364	3.5630
	TIFUKNN	0.1763	0.1683	0.5729	1.6677	1.8252	0.9255	0.6236	2.8502	2.5525	3.9346
	DNNTSP	$0.0871_{(0.0019)}$	$0.0792_{(0.0016)}$	$0.4303_{(0.0048)}$	$0.5513_{(0.0268)}$	$1.7653_{(0.0506)}$	$0.1037_{(0.0363)}$	$0.2545_{(0.0126)}$	$2.7432_{(0.0130)}$	$2.0765_{(0.0255)}$	$2.8581_{(0.0263)}$
	ReCANet	$0.1730_{(0.0011)}$	$0.1625_{(0.0010)}$	$0.5655_{(0.0018)}$	$1.0334_{(0.0204)}$	$1.6296_{(0.0334)}$	$0.4417_{(0.0127)}$	$0.4775_{(0.0075)}$	$2.7926_{(0.0066)}$	$2.3489_{(0.0139)}$	$3.3279_{(0.0197)}$
TaFeng	G-TopFreq	0.0812	0.0893	0.2565	5.4916	2.4948	2.9267	1.0932	3.1283	3.3155	7.6939
	Dream	$0.0888_{(0.0032)}$	$0.0924_{(0.0011)}$	$0.2798_{(0.0059)}$	5.4916	$2.5259_{(0.0072)}$	$0.6672_{(0.0055)}$	$1.0929_{(0.0001)}$	$3.1280_{(0.0001)}$	$3.3150_{(0.0002)}$	7.6939
	P-TopFreq	0.1087	0.0983	0.3608	0.4059	0.8820	5.5137	0.2735	2.7269	2.0944	2.4192
	GP-TopFreq	0.1183	0.1018	0.3740	0.4357	0.8986	5.6991	0.2775	2.7335	2.1050	2.4492
	UP-CF@r	0.1406	0.1222	0.4325	1.6427	1.4053	4.7278	0.6896	2.9401	2.7237	3.6597
	TIFUKNN	0.1618	0.1419	0.4697	1.1420	1.3862	0.8630	0.6299	2.8835	2.6074	3.1573
	DNNTSP	$0.1483_{(0.0004)}$	$0.1239_{(0.0004)}$	$0.4482_{(0.0011)}$	$1.4265_{(0.0645)}$	$1.5422_{(0.0444)}$	$0.3508_{(0.0332)}$	$0.4712_{(0.0246)}$	$2.8531_{(0.0121)}$	$2.4183_{(0.0367)}$	$3.4427_{(0.0647)}$
	ReCANet	$0.1287_{(0.0002)}$	$0.1188_{(0.0002)}$	$0.4195_{(0.0005)}$	$0.9450_{(0.0045)}$	$1.2147_{(0.0080)}$	$0.1644_{(0.0098)}$	$0.4343_{(0.0072)}$	$2.8093_{(0.0042)}$	$2.3376_{(0.0114)}$	$2.9597_{(0.0045)}$

↑ indicates that higher scores are better; ↓ indicates that lower scores are better; ○ means that the closer the value is to 0, the better the fairness. Numbers indicate the mean; subscripts in parentheses are the standard deviation; missing standard deviations are 0.0000.

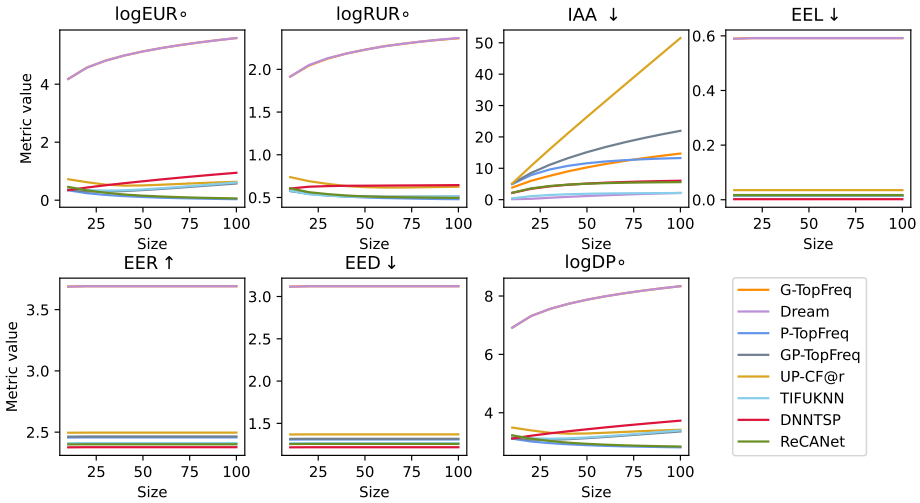


Figure 2.1: Performance of different NBR methods in terms of item fairness with varying basket size. Each plot represents one item fairness metric.

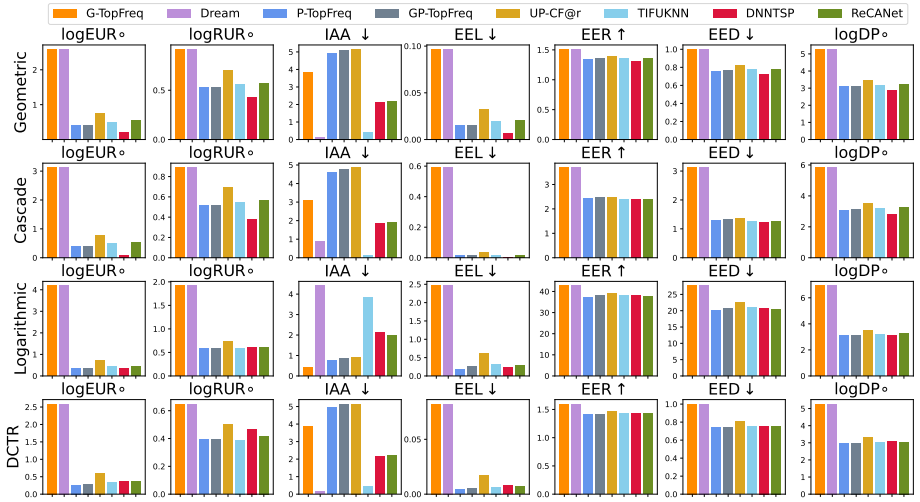
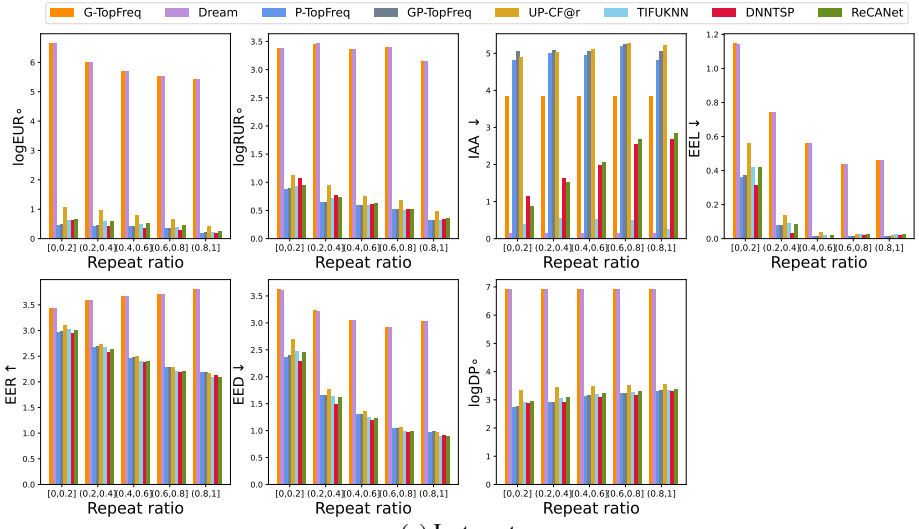


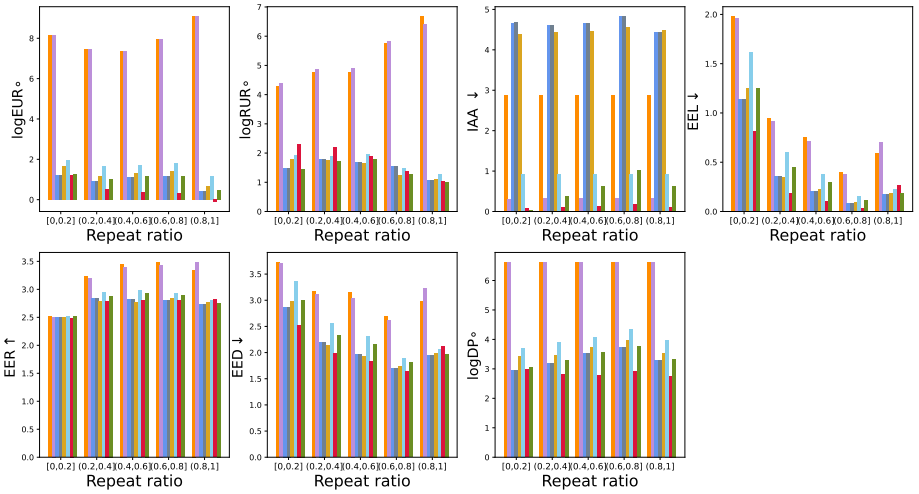
Figure 2.2: Item fairness with varying position weighting models. Each column represents an item fairness metric. Each row represents a position weighting model.

ordering adjustments happen among methods with similar fairness values. For example, for logEUR in Figure 2.2, the relative fairness order of TIFUKNN and DNNTSP is changed when the position weight model is changed from Logarithmic to DCTR. (iii) We find that the exposure values for each position weighting model is quite different. Since these metrics capture the gap between exposure and relevance, this could explain why their values change significantly in Figure 2.2 when using different position weighting models.

Repeat purchase behavior. To answer **RQ1.4**, we group users into five subgroups based



(a) Instacart



(b) Dunnhumby

Figure 2.3: Item fairness under different repeat ratios.

on their repeat consumption behavior. Here, *repeat ratio* is defined as the proportion of repeat items in the ground-truth basket for each user [63]. Hence, we create five subgroups of users with a repeat ratio of $[0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$, and $(0.8, 1.0]$, respectively. We calculate fairness metrics for each subgroup to test the sensitivity of fairness metrics to different degrees of repeat purchase behavior and report the results in Figure 2.3. We can observe that: (i) The fairness metrics show clear differences for different user groups, indicating that repeat purchase behavior does affect item fairness. (ii) The pattern on the Instacart dataset is relatively clear (Figure 2.3a). logEUR, logRUR, and EEL consider user groups with higher repeat ratios fairer because they depend highly on the utility of the NBR methods. As NBR

utility increases for higher repeat ratios [63], fairness for higher repeat ratios increases too. (iii) EER also belongs to the notion of equal opportunity, but these repeat-biased methods become more unfair as the repeat ratio increases, which is inconsistent with logEUR, logRUR, and EEL. (iv) IAA is only highly related to the predicted relevance, which is not affected by the utility of NBR methods. IAA values for explore-biased methods (GTopFreq and Dream) are stable. Only DNNTSP and ReCANet become more unfair when the repeat ratio increases, while IAA values for other methods fluctuate. (v) EED and logDP only measure whether the distribution of the popular and unpopular groups is uniform. From logDP, we see that, except for GTopFreq and Dream, the popular group gains more exposure when the repeat ratio increases. Since GTopFreq and Dream only recommend popular items to users, their logDP measurements do not change. However, EED indicates fairer for all methods as the repeat ratio goes up, which is contradictory to logDP. The position weighting model of EED is cascade, giving a discount to the exposure of correctly predicted items. User groups with higher repeat ratios have more correctly predicted items; therefore, the EED values decrease even though there are actually more spots given to popular items for user groups with higher repeat ratios. (vi) Patterns on the Dunnhumby (Figure 2.3b) and TaFeng datasets[7] are more complex. Some metrics do not change monotonically as the repeat ratio goes up. For instance, EEL and EED agree that the recommendation for user group with repeat ratio $(0.6, 0.8]$ demonstrates the best item fairness.

2.5 Conclusion

In this chapter, we have reproduced the fairness metrics implementation and empirical experiments in [98] in the context of NBR to answer RQ1:

> *What is the item fairness performance of current next basket recommendation methods?*

Specifically, we measure the item fairness of NBR methods and find that these metrics give different fairness rankings of the NBR methods. However, most of the metrics agree that repeat-biased methods are fairer than explore-biased methods. Different from the observations in [98], IAA is the most sensitive metric to both basket size and position weighting models. Finally, we analyze how repeat purchase behavior affects item fairness from the perspective of both equal opportunity and statistical parity. Above all, we recommend using EEL and EED for NBR since they show high robustness towards parameter configuration and various position weighting models, and can measure fairness for multiple groups.

Implications. Our work confirms that fairness metrics show different patterns of sensitivity for different information access systems due to the characteristics of the scenarios and subtle differences in metrics implementation. For the sake of rigor, we suggest testing the sensitivity of fairness metrics for every specific scenario, following the evaluation framework used by us and by Raj and Ekstrand [98] to ensure the

[7]The pattern on the TaFeng dataset is similar to that on the Dunnhumby dataset and we report the results on the TaFeng dataset in the Appendix of this chapter.

employed metrics are reliable. In addition, repeat-biased NBR methods exhibit better statistical parity fairness, because users' personalized repeat purchase behavior helps mitigate the popularity bias commonly observed in recommender systems; they also achieve better equal opportunity fairness owing to their higher recommendation accuracy. However, we cannot conclude that repeat-biased NBR is therefore preferable, since repeat bias itself is a serious problem: it reduces users' chances of discovering novel items. Thus, although repeat-biased NBR methods appear fairer, we still need to mitigate repeat bias, which we address in Chapter 4.

Limitations and future work. This work is the first attempt to study item exposure fairness in the NBR domain, which provides a reference for the optimization direction of subsequent NBR methods. We have mainly considered group fairness grouped by popularity, but it is worth examining other ways of grouping, such as by brand or category. From a practical perspective, equal opportunity is a safe and reasonable optimization goal, however, we cannot conclude whether statistical parity is applicable to NBR since it is not reasonable to assign equal exposure to popular and unpopular groups in practical grocery shopping scenarios. Nevertheless, future work should investigate the ideal exposure distribution of the two groups.

In the next chapter, we build on the fairness insights gained in this chapter to develop an NBR method that jointly optimizes for accuracy and beyond-accuracy objectives, extending our scope beyond fairness to include diversity.

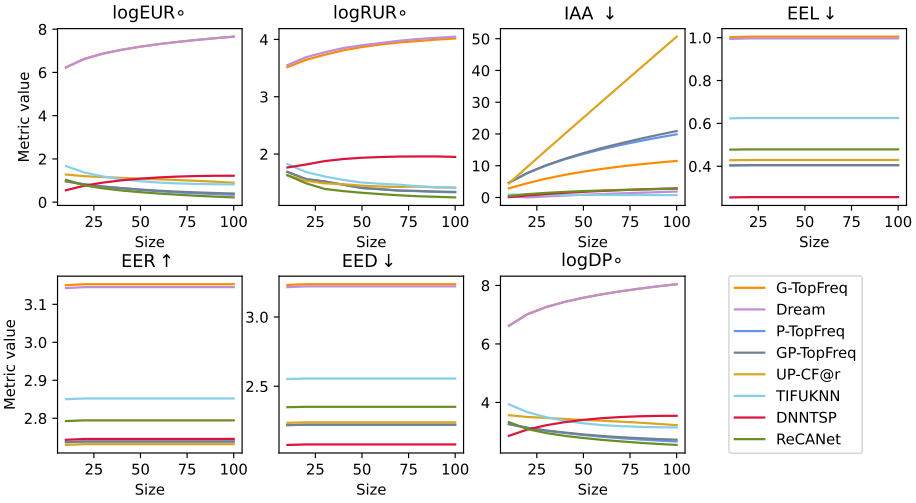
2.6 Appendix

2.6.1 Reproducibility

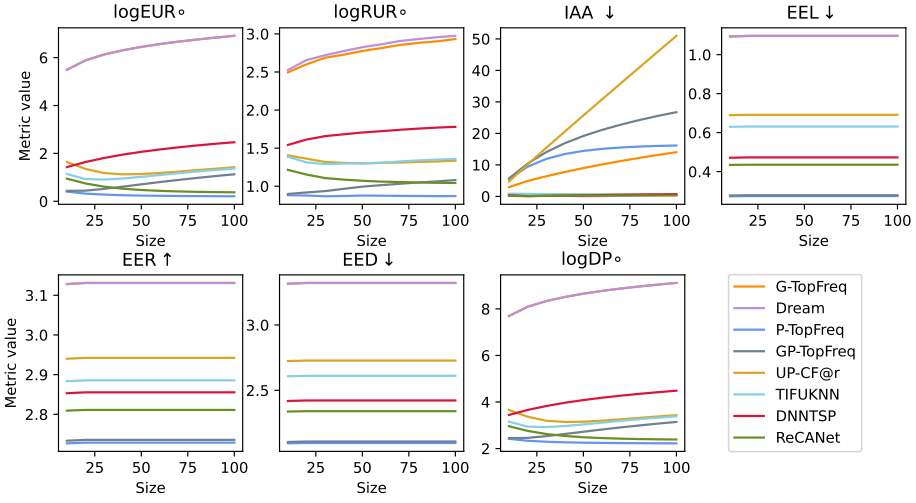
We release our code and hyperparameters at: https://github.com/lynEcho/NBR-fairness.

2.6.2 Additional plots

Figure 2.4 shows the variation of fairness values as basket size changes on Dunnhumby and TaFeng. Figure 2.5 shows the variation of fairness values as position weighting models change on Dunnhumby and TaFeng. Figure 2.6 shows the fairness metrics for different user groups on TaFeng.

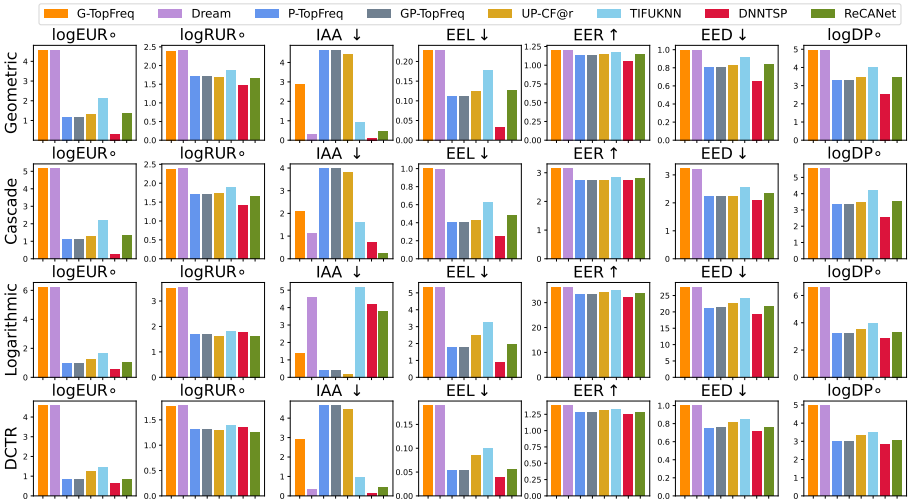


(a) Dunnhumby

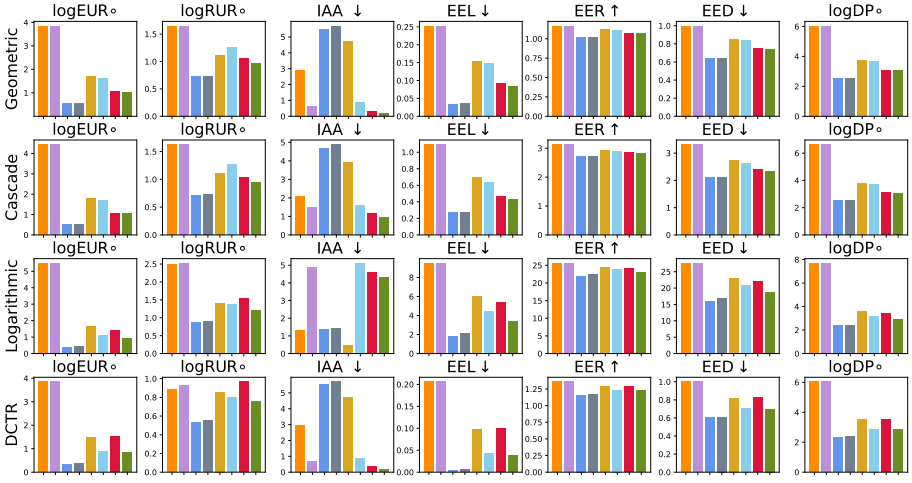


(b) TaFeng

Figure 2.4: Performance of different NBR models in terms of item fairness with varying basket size



(a) Dunnhumby



(b) TaFeng

Figure 2.5: Item fairness with varying position weighting models

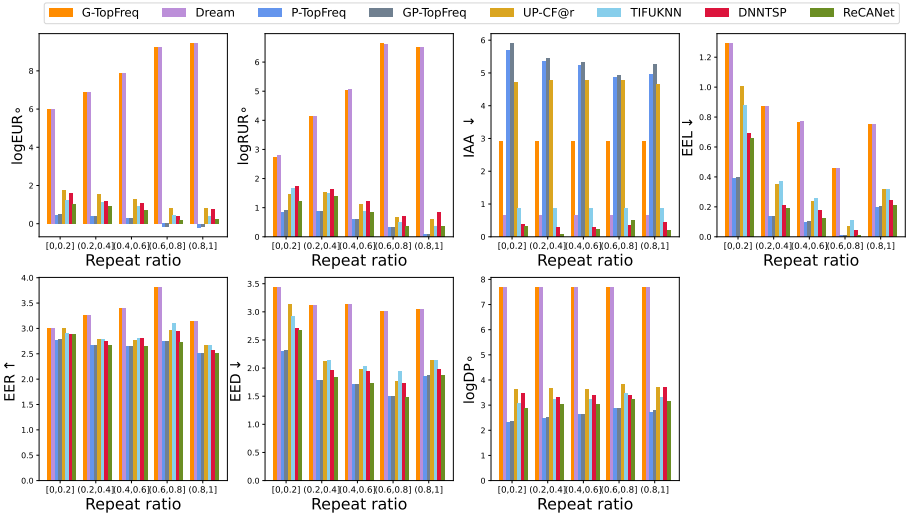


Figure 2.6: Item fairness under different repeat ratios on TaFeng

3

Are We Really Achieving Better Beyond-Accuracy Performance in Next Basket Recommendation?

As we explained in Chapter 2, next basket recommendation (NBR) is a special type of sequential recommendation that is increasingly receiving attention. So far, most NBR studies have focused on optimizing the accuracy of the recommendation, whereas optimizing for beyond-accuracy metrics, e.g., item fairness and diversity remains largely unexplored. Recent studies into NBR have found a substantial performance difference between recommending repeat items and explore items. Repeat items contribute most of the users' perceived accuracy compared with explore items [63]. These findings motivate us to investigate:

RQ2 *How can we design a basket recommendation method that jointly targets accuracy and beyond-accuracy objectives, and does the current evaluation paradigm adequately support this goal?*

3.1 Introduction

Recommender systems have become an essential instrument for connecting people to the content, services, and products they need. In e-commerce, more and more consumers purchase food and household products online instead of visiting physical retail stores [56]. The COVID-19 pandemic has only accelerated this shift [90]. In this scenario, consumers usually purchase a *set* of items at the same time, a so-called *basket*. *Next basket recommendation* (NBR) is a type of sequential recommendation that caters to this scenario: baskets are the target of recommendation and historical sequential data consists of users' interactions with baskets. NBR has increasingly been attracting attention in recent years [4]. Many methods, based on different machine learning techniques, have been proposed for accurate recommendations, e.g., Markov chain

This chapter was published as M. Li, Y. Liu, S. Jullien, M. Ariannezhad, A. Yates, M. Aliannejadi, and M. de Rijke. Are we really achieving better beyond-accuracy performance in next basket recommendation? In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 924–934, 2024.

(MC)-based methods [101, 116], frequency and nearest neighbor-based methods [33, 45], RNN-based methods [44, 57, 96, 137], and self-attention methods [21, 107, 138].

Repetition vs. exploration in NBR. Recently, Li et al. [63] have assessed the performance of state-of-the-art NBR in terms of repeat and explore items: items that a user has interacted with before and items that they have never interacted with before, respectively. The authors distinguish between the task of repetition recommendation (recommending *repeat items*) and the task of exploration recommendation (recommending *explore items*). Repetition and exploration recommendations have different levels of difficulty, where recommending items that are regularly present in a user's baskets is shown to be a far easier task [63]. Building on these findings, repetition-only [3, 54] and exploration-only [61] methods have been proposed to optimize the accuracy of next basket recommendation.

Accuracy and beyond-accuracy metrics. Even though accuracy naturally serves as the most important objective of recommendations, it is widely recognized that it should not be the sole focus. Beyond-accuracy metrics such as item fairness [30, 40, 123, 125] and diversity [20, 146, 149] also play crucial roles in evaluating recommendation services. Such beyond-accuracy metrics have gained increasing attention and have been optimized in a range of recommendation scenarios [136, 149]. In the NBR scenario, however, beyond-accuracy metrics have been far less studied than accuracy-based metrics. In this chapter, we help to address this knowledge gap. Following the paradigm of multiple-objective recommender systems [49], it is widely recognized that there is a trade-off between accuracy and beyond-accuracy metrics. E.g., diversity goals are reckoned to stand in contrast with accuracy. Put differently, a method achieving a better beyond-accuracy performance while maintaining the same level of accuracy performance is considered to be a success [136, 149]. How can we achieve a reasonable balance between accuracy and beyond-accuracy metrics in NBR?

Potential "short-cuts" to balancing accuracy and beyond-accuracy metrics. Besides the imbalance between repetition and exploration [60, 62–64], Li et al. also found that repeat items contribute most of the accuracy, whereas the explore items in the recommended basket contribute very little to the user's perceived utility. As Table 3.1 summarizes, there are essential differences between the repetition and exploration tasks, which explain the substantial performance differences between the two tasks.

Inspired by these findings, we hypothesize that there may be a "short-cut" strategy to optimize for both accuracy and beyond-accuracy metrics, which contains two aspects: (i) *accuracy*: Predict repeat items to achieve good accuracy: predicting repeat items is much easier than predicting explore items [63], and (ii) *beyond-accuracy*: Use explore items to improve beyond-accuracy metrics: it is very difficult to recommend quality explore items. Thus, exchange the low accuracy that is typically achieved on such items for beyond-accuracy metrics, i.e., trade accuracy for diversity and item fairness. We call this NBR strategy a short-cut strategy because it avoids making the fundamental trade-off between accuracy and beyond-accuracy metrics.

TREx framework. To operationalize our short-cut idea, and check whether the "short-cut" strategy can be made to work, we propose the *two-step repetition-exploration* (TREx) framework. TREx decouples the prediction of repeat items and explore items. Specifically, TREx uses separate models for predicting (a) repeat items, and (b) explore

Table 3.1: Comparison of the repetition and exploration tasks in NBR.

Aspect	**Repetition**	**Exploration**
Task difficulty	Easy	Difficult
Number of items	Dozens	Thousands
Item interactions	Previous	None
Users' interest	With feedback	Without feedback
Task type	Re-consume	Infer new

items, and then combines the outcomes of the two prediction models to generate the next basket. In contrast, existing NBR methods usually output the scores/probabilities of all items and then select the top-k items to fill up a basket to be recommended, ignoring the differences between repeat and explore items.

For TREx's repeat item prediction, we propose a simple yet effective probability-based method, which considers the item characteristics and users' repurchase frequency. For exploration recommendations, we design two strategies that cater to the different beyond-accuracy metrics. The flexibility of TREx allows us to design suitable models for repetition and exploration, with the possibility of controlling the proportions of repetition and exploration to investigate the relations between accuracy and various beyond-accuracy metrics.

Findings and reflections. We consider two types of widely-used beyond-accuracy metrics, i.e., diversity and item fairness. Specifically, we investigate five fairness metrics (i.e., logEUR, logRUR, EEL, EED, and logDP) [75, 98] and three diversity metrics (i.e., ILD, Entropy, and DS) [136]. To provide an overall understanding of these metrics, we group them according to different levels of connection with accuracy as follows: (i) Strong connection: logRUR, (ii) Weak connection: logEUR, EEL, EED (iii) No connection: logDP, ILD, Entropy, DS. Briefly, the strong connection between logRUR and accuracy stems from the fact that logRUR uses ground truth relevance to discount the exposure, making sure that only correctly predicted items contribute to effective exposure. The connection between logEUR, EEL, and accuracy is weak because they just ensure the exposure distribution across groups of recommended results is close to the group exposure distribution of ground truth, without considering whether the exposure is contributed by correctly predicted items. Since the position weighting model of EED considers ground truth, EED shows a weak connection. There is no connection between accuracy and logDP, ILD, Entropy, and DS because their exposure distributions across groups are designed to reflect a specific distribution. The strength of the connection between a beyond-accuracy metric and accuracy determines whether there is a short-cut towards optimizing both accuracy and the beyond-accuracy metric.

We perform experiments on two brick-and-mortar retailers' NBR datasets, considering six NBR baselines and eight metrics. The experimental results show that: (i) State-of-the-art accuracy can be achieved by only recommending repeat items via the proposed simple yet effective repetition model. (ii) Leveraging the "short-cut" using TREx achieves "better" beyond-accuracy performance w.r.t. seven out of eight

beyond-accuracy metrics. (iii) In terms of the item fairness metric having a strong connection with accuracy (i.e., logRUR), it is more difficult to achieve better beyond-accuracy metrics via the proposed strategy.

Stepping back. Instead of blindly claiming TREx with the designed modules as a state-of-the-art method for optimizing both accuracy and various beyond-accuracy metrics, we reflect on and challenge our evaluation paradigm in the definition of success in this setting. The core question is:

> *Are we really achieving better beyond-accuracy performance in next basket recommendation?*

Two perspectives offer different ways forward for researchers and practitioners to address this question:

(1) If we are willing to sacrifice the accuracy of the exploration, then superior beyond-accuracy performance can be achieved by leveraging the "short-cut" strategy via TREx, which is straightforward and efficient. This "short-cut" strategy must be considered before developing more sophisticated and elaborate approaches.

(2) Conversely, if we believe it is unreasonable to sacrifice the accuracy of exploration [122], the existence of the "short-cut" strategy reveals flaws in our current evaluation paradigm to demonstrate an NBR method's superiority. A fine-grained analysis (i.e., distinguishing between repetition and exploration) needs to be performed to check whether "better" beyond-accuracy is achieved by triggering the "short-cut" strategy, which would hurt the exploration accuracy after all.

Our contributions. The main contributions of the chapter are:

- We identify a "short-cut" strategy (i.e., sacrificing accuracy for exploration and using explore items to optimize for beyond-accuracy metrics), which could achieve "better" beyond-accuracy metrics without degrading accuracy.
- We propose a simple repetition recommendation model considering item features and users' repurchase frequency, which can achieve state-of-the-art NBR accuracy by only recommending repeat items.
- We propose TREx, a flexible two-step repetition-exploration framework for NBR, which allows us to control the trade-off between accuracy and beyond-accuracy metrics w.r.t. the recommended baskets.
- We conduct experiments on two datasets w.r.t. eight beyond-accuracy metrics, and find that leveraging "short-cuts" via TREx can achieve better performance on a wide range of metrics. We also find that the stronger the connection with accuracy, the more challenging it becomes to utilize a "short-cut" strategy to enhance a beyond-accuracy metric.
- We reflect on, and challenge, existing evaluation paradigms, and find that a fine-grained level analysis can provide a complementary view of a method's performance.

3.2 Related Work

We summarize related research on next basket recommendation and beyond-accuracy metrics.

Next basket recommendation. The NBR problem has been studied for many years. Factorizing personalized Markov chains (FPMC) [101] uses matrix factorization and Markov chains to model users' general interest and basket transition relations. HRM [116] applies aggregation operations to learn a hierarchical representation of baskets. RNNs have been adapted to the NBR task to learn long-term trends by modeling the whole basket sequence. E.g., Dream [137] uses max/avg pooling to encode baskets. Sets2Sets [44] adapts an attention mechanism and adds frequency information to improve performance. Some methods [57, 119] consider the underlying item relations to get a better representation. Yu et al. [138] argue that item-item relations between baskets are important, and use GNNs to use these relations. Some methods [5, 58, 107, 117] exploit auxiliary information, including product categories, amounts, prices, and explicit timestamps. TIFUKNN [45] and UP-CF@r [33], frequency-neighbor-based methods, model temporal patterns, and then combine these with neighbor information or user-wise collaborative filtering. Li et al. [63] provide several metrics to evaluate repetition and exploration performance in the NBR task and find that the repetition task is easier than the exploration task. Inspired by this analysis, repetition-only [3, 54] and exploration-only [61] models were proposed for next basket recommendation. Existing NBR work mainly focuses on optimizing accuracy whereas this chapter extends to various beyond-accuracy metrics for NBR.

Beyond-accuracy metrics. In addition to accuracy, there are various beyond-accuracy metrics (i.e., diversity, fairness, novelty, serendipity, coverage) we need to consider when making recommendations [30]. Diversity is a crucial factor in meeting the diverse demands of users [19, 97, 118, 146]. Recently, empirical and revisitation studies [79, 136] have been conducted to explore the trade-off between accuracy and diversity. The concepts of fairness and item exposure have emerged as crucial considerations since items and producers play pivotal roles within a recommender system and its ecosystem. Related metrics measure whether items receive a fair share of exposure according to different definitions of fairness. Current research on fairness primarily focuses on individual or group fairness, either from the customer's perspective, adopting a user-centered approach [10], or from the provider's viewpoint, adopting an item-centered approach [85, 140], or a two-sided approach [86, 123, 125]. Recently, Liu et al. [75] evaluated item fairness on existing NBR methods to investigate the robustness of different fairness metrics. Unlike the work listed above, this chapter is not limited to optimizing a specific type of metric. It examines the possibility of using a "short-cut" strategy to seemingly optimize various beyond-accuracy metrics and provides insights w.r.t. evaluation paradigms when extending NBR optimization and evaluation to these beyond-accuracy metrics.

3.3 Task Formulation and Definitions

We describe the next basket recommendation problem and formalize the notions of repetition and exploration. Our notation is summarized in Table 3.2.

Next basket recommendation. Given a set of users $U = \{u_1, u_2, \ldots, u_n\}$ and items $I = \{i_1, i_2, \ldots, i_m\}$, $S_u = \{B_u^1, B_u^2, \ldots, B_u^t\}$ represents the historical interaction sequence for user u, where B_u^t is the user's basket at time step t. B_u^t consists of a set of items $i \in I$, and the goal of the *next basket recommendation* task is to predict the following basket B_u^{t+1}, i.e., the items that the user would probably like, based on the user's past interactions S_u. We write $\hat{B}_u^{t+1}$ for the predicted basket and denote it by P_u for brevity:

$$P_u = \hat{B}_u^{t+1} = f(S_u), \tag{3.1}$$

where f is our basket generation algorithm. We assume that the user's attention and screen space is limited; hence, like previous studies [63, 75], we recommend fixed-size baskets of sizes 10 or 20.

Repetition and exploration. We assume that the set of items is fixed. Although this might not be the case in real-world settings, modeling the addition and deletion of items in the set of items is out of the scope of this chapter. With this assumption in mind, the addition of every new basket to the users' history, may translate into fewer items left to explore. To differentiate between the items coming from the exploration and repeat consumption behavior, for a user u and timestamp t, a set of items $I_{u,t}^{rep} \subset I$ are considered to be the "repeat items." The set of explore items $I_{u,t}^{expl}$ is simply its complement within the overall item set I. We define $I_{u,t}^{rep}$ as:

$$I_{u,t}^{rep} = I_{u,t-1}^{rep} \cup B_u^t. \tag{3.2}$$

This also means that $I_{u,1}^{rep} \subset \cdots \subset I_{u,t-1}^{rep} \subset I_{u,t}^{rep}$. Conversely, we have $I_{u,t}^{expl} \subset I_{u,t-1}^{expl} \subset \cdots \subset I_{u,1}^{expl}$.

The task of predicting the next basket for a user u is equivalent to predicting which items from $I_{u,t}^{rep}$ and $I_{u,t}^{expl}$ will appear in B_u^{t+1}. One way to solve this problem is to decouple it into two subtasks: the repetition subtask that aims to predict which items from $I_{u,t}^{rep}$ to recommend, and the exploration task that recommends items from $I_{u,t}^{expl}$. Table 3.1 shows the different characteristics w.r.t. the repetition and exploration tasks.

3.4 Evaluation Metrics

Next, we describe the accuracy and beyond-accuracy metrics (i.e., fairness and diversity) considered in the chapter.[1]

Accuracy. In terms of accuracy, we use three metrics that are widely used for the NBR task: $Recall@k$, $NDCG@k$, and $PHR@k$. $Recall$ measures the ability to find all items that the user will purchase in the next basket; *NDCG* is a ranking metric that also

[1] We only provide brief introductions of each metric; more detailed information (e.g., function, responsibility, etc.) can be found in the original papers and relevant survey papers [75, 98, 149].

Table 3.2: Notation used in the chapter; fairness related notation is adapted from [75, 98].

Symbol	Description
$u \in U$	Users
$i \in I$	Items
S_u	Sequence of historical baskets for u
B_u^t	t-th basket in S_u, a set of items $i \in I$
$I_{u,t}^{rep}$	Set of repeat items for u up to timestamp t
$I_{u,t}^{expl}$	Set of explore items for u up to timestamp t
T_u	Ground-truth basket for u that we aim to predict
T_u^{rep}	Set of *repeat items* in the ground truth basket T_u for u
T_u^{expl}	Set of *explore items* in the ground truth basket T_u for u
P_u	Predicted basket for u
P_u^{rep}	Set of *repeat items* in the predicted basket P_u for u
P_u^{expl}	Set of *explore items* in the predicted basket P_u for u
$G(P)$	Group alignment matrix for items in P
G^+	Popular group
G^-	Unpopular group
$\mathbf{a}_P$	Exposure vector for items in P
ϵ_P	The exposure of groups in P $(G(P)^T \mathbf{a}_P)$

considers the order of the items; *PHR* is a user level measurement which represents the ratio of users whose recommended basket contains the item in the ground-truth.

Fairness. Assume $\pi(P \mid u)$ is a user-dependent distribution and $\rho(u)$ is a distribution over users; overall, the recommended item rankings among all users follow the following distribution: $\rho(u)\pi(P \mid u)$. $\epsilon_P = G(P)^{\mathrm{T}}\mathbf{a}_P$ is the group exposure within a recommended basket.[2] Its expected value $\epsilon_\pi = E_{\pi\rho}[\epsilon_P]$ is the group exposure among all the recommended baskets. Following [75, 98], we select a set of well-known fairness metrics and cover two types of fairness considerations as follows:[3]

(1) *Equal opportunity.* Promote equal treatment based on merit or utility, regardless of group membership [75, 98]. (i) *Exposed utility ratio (EUR)* [104] quantifies the deviation from the objective that the exposure of each group is proportional to its utility $Y(G)$. (ii) *Realized utility ratio (RUR)* [104] models actual user engagement, the click-through rates for the groups $\Gamma(G)$ are proportional to their utility. (iii) *Expected exposure loss (EEL)* [28] is the distance between expected exposure and target exposure ϵ^*, which is the exposure under the ideal policy.

(2) *Statistical parity.* Ensure comparable exposure among groups. (i) *Expected exposure disparity (EED)* [28] measures the inequality in exposure distribution across groups. (ii) *Demographic parity (DP)* [104] measures the ratio of average exposure given to the two groups. Following [98], we reformulate DP as logDP to tackle the issue of empty-group scenarios and improve interpretability. Exposed utility ratio (logEUR)

[2]The formula to compute the exposure vector $\mathbf{a}_P$ using different position weighting models can be found in [75, 98].

[3]The item fairness metric Inequity of Amortized Attention [9] is not used in this chapter since some baselines do not have predicted relevance for items.

Table 3.3: Summary of fairness and diversity metrics; fairness metrics are adapted from [98]. ↑ indicates that higher values are better; ↓ indicates that lower values are better; ○ means that the closer the value is to 0, the better the performance.

Category	Metrics	Goal	Better	Accuracy connection
Equal opportunity	logRUR	Click-through rate proportional to relevance	○	Strong
	logEUR	Exposure proportional to relevance	○	Weak
	EEL	Exposure matches ideal (from relevance)	↓	Weak
Statistical parity	EED	Exposure well-distributed	↓	Weak
	logDP	Exposure equal across groups	○	None
Diversity	ILD	Average distance between categories for each pair of items in the list	↑	None
	Entropy	Entropy of item category distribution in the list	↑	None
	DS	Number of categories divided by the number of items in the list	↑	None

and realized utility ratio (logRUR) are defined in a similar manner.

Diversity. Following [136], we consider the following widely-used diversity metrics, which satisfy users' diversified demands. (i) *Intra-list diversity (ILD)* [12, 19] measures the average distance between every pair of items in the recommendation list (P_u), where d_{ij} is the Euclidean distance between the respective embeddings of categories (ii) *Entropy* [118, 151] quantifies the dispersion of item category distribution in the recommendation list P_u; a higher degree of dispersion in the category distribution corresponds to increased diversity. (iii) *Diversity score (DS)* [71] is calculated as the number of interacted/recommended categories divided by the number of interacted/recommended items. As shown in Table 3.3, we can group beyond-accuracy metrics according to their connection with accuracy.

3.5 A Two-Step Repetition-Exploration Framework

Given the differences depicted in Table 3.1, we propose a *two-step repetition-exploration* (TREx) framework for NBR. TREx assembles recommendations from a repetition module and an exploration module. TREx allows one to easily swap out the sub-algorithms used for repetition and exploration. In the first step, we model the repetition and exploration behavior separately to get candidates from both sources. Then, we generate the recommended basket from those candidates in the second step. The main architectural differences between previous approaches to the NBR problem, which typically consists of a single treatment of all items, and TREx, which treats repeat and

Algorithm 1: TREx Framework

Data: Basket sequence S, basket size k, repetition confidence threshold v
Result: Recommended basket B_u^{t+1} for each user u,
1 Calculate the repetition feature $RepI(i)$ for each item;
2 **for** *each user* u **do**
3 Get repeat items $I_{u,t}^{rep}$, and explore items $I_{u,t}^{expl}$;
4 Calculate the repetition score $RepS^u(i)$ for each $i \in I_{u,t}^{rep}$;
5 Remove items i from $I_{u,t}^{rep}$, when $RepS^u(I) < v$;
6 Rank $I_{u,t}^{rep}$ according to $RepS^u(i)$ in descending order;
7 Initialize next basket B_u^{t+1};
8 **if** $|I_{u,t}^{rep}| < k$ **then**
9 Fill B_u^{t+1} using $I_{u,t}^{rep}$;
10 m $\leftarrow k - |I_{u,t}^{rep}|$;
11 Fill m empty slots of B_u^{t+1} using explore items via exploration module;
12 **else**
13 Fill B_u^{t+1} using top-k of $I_{u,t}^{rep}$;
14 **end**
15 **end**

explore items differently. The pseudo-code for TREx is given in Algorithm 1. Next, we describe the three modules that make up TREx.[4]

3.5.1 Repetition module

As the repetition task is a much simpler task than exploration, we therefore design a repetition module targeted at improving the accuracy. Intuitively, if a user consumed an item several times in the past, they are likely to repurchase that item in the next basket. Thus, frequency information is a strong signal for repetition prediction [113]. The personal item frequency (PIF) introduced in TIFUKNN [45] and the recency window in UP-CF@r [33] both capture temporal dependencies by focusing more on recent behavior. However, they do not capture the item characteristics w.r.t. repurchasing. For example, a purchase of a bottle of milk and a pan is more likely to be followed by a repurchase of milk rather than a pan, even if both currently have the same purchase frequency. To consider both item features and user interest simultaneously, we use the repetition score $RepS^u(i)$ to represent the repurchase score of item i for user u. This score is decomposed into two parts, the item-specific repurchase feature $RepI(i)$ and the user's interest E_i^u in item i. Formally:

$$RepS^u(i) = E_i^u \cdot RepI(i) \,. \tag{3.3}$$

This corresponds to line 4 in Algorithm 1. Given the items in the dataset $I = \{i_1, i_2, \ldots, i_m\}$, we need to derive the repurchase feature $RepI(i)$ for each item in

[4]Theoretically, TREx allows us to choose or design the suitable repetition and exploration modules both targeted at the accuracy to achieve state-of-the-art performance. However, we aim to investigate the "short-cut" and relationship between accuracy and various beyond-accuracy metrics.

the training set. First, the repurchase frequency $Rep^F(i)$ can be calculated by gathering the statistical information across users. To mitigate the impact of abnormally high values in some users, we introduce a hyperparameter α to discount the repurchase frequency of item i:

$$Rep^F(i) = \frac{\sum_U (\text{item } i \text{ repurchase frequency})^\alpha}{\#\text{users who bought item } i \text{ at least once}} . \tag{3.4}$$

In addition, some items might only have a few samples, which might lead to low confidence about their repetition feature estimation. We leverage the average estimate $\overline{RepF}$ across all items as supplementary information to help items with a few samples. Then, the final repetition feature is given by:

$$RepI(i) = Rep^F(i) + \frac{\overline{RepF}}{N_i}, \tag{3.5}$$

where N_i is the number of users who bought item i. Thus, the average $\overline{RepF}$ will have a small effect on $RepI(i)$ when we have more samples to compute item-specific features. This corresponds to line 1 in Algorithm 1.

The item frequency in a user's historical baskets can partially reflect the user's interest. Yet, user interests can also be dynamic. To model temporal dependencies, we introduce a time-decay factor β, which makes the recent interactions have more impact on the interest E_i^u. Assume that a specific item i was purchased by the user u several times in their historical baskets $\{B_u^{l_1}, B_u^{l_2}, \ldots, B_u^{l_m}\}$; the corresponding position set is denoted as $L_i = \{l_1, l_2, \ldots, l_m\}$; then E_i^u is defined as:

$$E_i^u = \sum_{j=1}^{m} \beta^{T-l_j} , \tag{3.6}$$

where T represents the length of the user's basket. TREx's repeat recommendation model takes item features, user interests, and the temporal order of baskets into consideration. We treat the items in baskets independently and calculate the repetition score $RepS$ for all items that appeared in the previous baskets for each user, which will be used in the final basket generation process.

3.5.2 Exploration module

As it is more challenging than repetition, exploration is also an important aspect of NBR. To complement the repetition module, we design different exploration modules, targeting item fairness and diversity, respectively. For each user u, the exploration candidates $I_{u,t}^{expl}$ are the set of items that the user never bought before.

Item fairness. According to [63], we find that NBR methods usually have varying degrees of popularity bias, which means they recommend more popular items compared to the ground truth and harm item fairness. Thus, we recommend unpopular items $i \in G^-$ for the exploration module for the sake of approaching the distribution of ground truth and decreasing the exposure gap between the popular and the unpopular groups. Specifically, we randomly sample explore items based on a sampling probability, which is calculated from the purchase frequency of unpopular items.

Diversity. Diversity optimizes for more dispersed categories in the predicted basket. For each user, we record categories of repetition candidates, rank exploration candidates

Table 3.4: Statistics of the processed datasets.

Dataset	#items	#users	Avg. basket size	Avg. #baskets per user	Repeat ratio	Explore ratio
Instacart	29,398	19,210	10.06	15.91	0.60	0.40
Dunnhumby	37,161	2,482	10.07	43.17	0.43	0.57

according to their popularity, and select explore items to fill in the B_u^{t+1} in turn. The category of each explore item differs from the categories already in B_u^{t+1}.

3.5.3 Basket generation module

To construct the final basket to be recommended by TREx for the accuracy objective, we adopt a repetition greedy approach and first consider the item candidates generated by the repetition module and fill the remaining slots via the exploration module. TREx$_{Fairness}$ and TREx$_{diversity}$ denote TREx with the exploration module targeted at fairness and diversity, respectively. For a user u, we get their repetition score $RepS^u(i)$, where $i \in I_{u,t}^{rep}$ (Algorithm 1, lines 3–4). First, we define a confidence threshold v for the repetition score and repetition items are removed from the $i \in I_{u,t}^{rep}$ when the corresponding $RepS^u(i) < v$ (line 5).[5] Then, $I_{u,t}^{rep}$ can be seen as the repetition candidates set. If the number of repetition candidates exceeds the basket size, the items with a high score will have priority to fill the basket (Algorithm 1, line 13). If the number of repetition candidates is smaller than the basket size, the basket is first filled with all items in the repetition candidates set $I_{u,t}^{rep}$. Then, we fill up the basket using the explore items via the exploration module, where m represents the number of empty slots (lines 9–11).

3.6 Experiments

3.6.1 Experimental setup

Datasets. We conduct experiments on two widely-used datasets: (i) Instacart,[6] which includes a large number of grocery orders from users; following [75, 87], ∼20000 users are randomly selected to conduct experiments; and (ii) Dunnhumby,[7] which contains two years' household-level transactions of 2500 frequent shoppers at a retailer. Following [3, 75], we sample users who have at least three baskets and remove items that appeared less than five times. The two datasets vary in the repeat ratio, i.e., the proportion of repeat items in the ground-truth baskets [63]. We focus on the fixed size (10 or 20) NBR problem. The statistics of the processed datasets are shown in Table 3.4. In our experiments, each dataset is partitioned according to [3, 33, 75, 87]. The training

[5]The confidence threshold v controls the proportion of repeat items and explore items in the recommendation, as well as the accuracy and beyond-accuracy trade-off in this chapter. We sweep repetition confidence bound v to get TREx variants with different accuracy and beyond-accuracy metrics performance.

[6]https://www.kaggle.com/datasets/yasserh/instacart-online-grocery-basket-analysis-dataset

[7]https://www.dunnhumby.com/source-files/

baskets encompass all user baskets except the last one. In cases where users have over 50 baskets in the training data, only their last 50 baskets are considered for inclusion in the training set. The final baskets of all users are then divided equally between a 50% validation set and a 50% test set. Figure 3.1 shows the distribution of users across repeat ratios, which is the proportion of *repeat items* in the ground-truth basket.

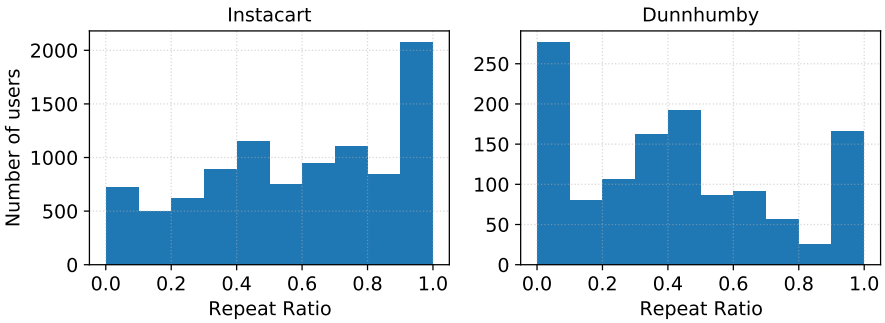


Figure 3.1: Distribution of users across different repeat ratios for Instacart and Dunnhumby.

NBR baselines. We compare TREx with 8 representative baselines, which we select based on their characteristics in the analysis performed in [63, 75], divided into three groups:

Simple baselines. (i) **G-TopFreq** uses the k most popular items in the dataset to form the recommended next basket. (ii) **P-TopFreq** is a personalized TopFreq method, which treats the most frequent k items in historical records of the user as the next basket. (iii) **GP-TopFreq** [63] is a simple combination of P-TopFreq and G-TopFreq, which first use P-TopFreq to fill the basket, then use G-TopFreq to fill the remaining slots.

Nearest neighbor-based methods. (i) **TIFUKNN** [45] is a state-of-the-art method that models the temporal dynamics of frequency information of users' past baskets to introduce personal item frequency (PIF), then it uses KNN-based method on the PIF. (ii) **UP-CF@r** [33] is a combination of recency aware user-wise popularity and user-wise collaborative filtering.

Neural network-based methods. (i) **Dream** [137] models users' global sequential basket behavior for NBR using recurrent neural network (RNN). (ii) **DNNTSP** [138] is a state-of-the-art method that leverages a GNN and self-attention techniques. It encodes item-item relations via a graph and employs a self-attention mechanism to capture temporal dependencies of users' basket sequences. (iii) **ReCANet** [3] is a repeat-only model for NBR, which uses user-item representations with historical consumption patterns via RNN.

Configurations. To assess group fairness (Section 3.4), we follow configurations from previous research [69, 75]; the group of items is determined by their popularity (i.e., the number of purchases recorded in the historical baskets of the dataset). The top 20% of items with the highest purchase frequency as the popular group (G^+), while the remaining 80% of items are assigned to the unpopular group (G^-). For the baseline methods, a

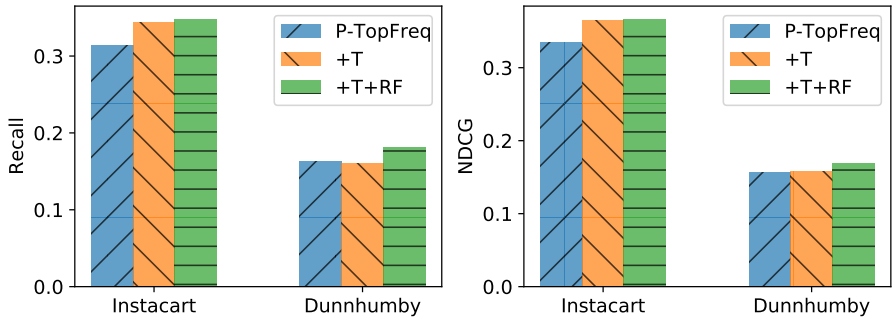


Figure 3.2: Performance of TREx-Rep when we add a time-decay factor β (+T), add both β and item-specific repetition feature $RepI(i)$ (+T+RF).

grid search is performed to find the optimal hyperparameters via the validation set. For TIFUKNN, the number of neighbors k is tuned on $\{100, 300, 500, 900, 1100, 1300\}$, the number of groups m is tuned on $\{3, 7, 11, 15, 19, 23\}$, the within-basket time-decayed ratio r_b and the group time-decayed ratio r_g are selected from $\{0.1, 0.2, \ldots, 0.9, 1\}$, and the fusion weight α is selected from $\{0, 0.1, \ldots, 0.9, 1\}$. For UP-CF@r, recency window r is tuned on $\{1, 5, 10, 25, 100, \infty\}$, locality q is tuned on $[1, 5, 10, 50, 100, 1000]$, and asymmetry α is tuned on $\{0, 0.25, 0.5, 0.75, 1\}$. For Dream, DNNTSP, and ReCANet, the item and user embedding size is tuned on $\{16, 32, 64, 128\}$. As to TREx, for the repetition module, α is selected from $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, and the time-decay factor β is selected from $\{0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0\}$.

3.6.2 Overall accuracy performance

By decoupling the repetition and exploration tasks, TREx-Rep optimizes for the repeat items prediction and accounts for the accuracy of the NBR performance. Table 3.5 shows the experimental results for TREx-Rep and the baselines. We observe that TREx-Rep surpasses two complex deep-learning-based methods (i.e., Dream and DNNTSP) by a large margin on the Dunnhumby and Instacart datasets, and TREx-Rep always achieves or matches the SOTA accuracy on both datasets across different accuracy metrics. Note that TREx-Rep achieves a competitive accuracy performance by only using part of the available slots in the basket.[8] Compared to the deep learning methods with complex architectures that try to learn basket representations and model temporal relations, TREx-Rep is very efficient due to its simplicity.

To investigate the effect of the repetition features and the improvement in repetition performance in NBR. We conduct experiments on TREx-Rep by gradually adding the time-decay factor β and item-specific repetition feature $RepI(i)$. The results are shown in Figure 3.2. The accuracy increases when we gradually integrate different factors into TREx-Rep, which indicates that both the time-decay factor β and the

[8] As TREx-Rep only recommends repeat items, the basket could not be fulfilled when the number of user's repeat items (historical items) is smaller than the basket size. ReCANet also only recommends repeat items, however, it is a complex neural-based model, which is much slower than the proposed TREx-Rep module.

Table 3.5: Comparison of TREx-Rep (repetition-module only) against baselines and two types of state-of-the-art methods; boldface indicates the maximum; underlining indicates the second best performing method. † indicates that TREx-Rep results achieve the same level of performance as SOTA baselines (paired t-test).

Dataset	Metric	G-TopFreq	P-TopFreq	GP-TopFreq	UP-CF@r	TIFUKNN	Dream	DNNTSP	ReCANet	TREx-Rep
Instacart	Recall@10	0.0704	0.3143	0.3150	0.3377	0.3456	0.0704	0.3295	**0.3490**	0.3476†
	NDCG@10	0.0817	0.3339	0.3343	0.3582	0.3657	0.0817	0.3434	**0.3699**	0.3661†
	PHR@10	0.4600	0.8447	0.8460	0.8586	0.8639	0.4600	0.8581	**0.8668**	0.8655†
	Recall@20	0.0973	0.4138	0.4168	0.4405	0.4559	0.0979	0.4339	**0.4562**	0.4557†
	NDCG@20	0.0962	0.3889	0.3902	0.4161	0.4271	0.0968	0.4018	**0.4303**	0.4269†
	PHR@20	0.5302	0.8921	0.8959	0.9045	**0.9098**	0.5346	0.9033	0.9097	0.9092†
Dunnhumby	Recall@10	0.0897	0.1628	0.1628	0.1699	0.1763	0.0896	0.0871	0.1730	**0.1815**†
	NDCG@10	0.0798	0.1562	0.1562	0.1639	0.1683	0.0759	0.0792	0.1625	**0.1689**†
	PHR@10	0.3795	0.5399	0.5399	0.5536	0.5729	0.3873	0.4303	0.5655	**0.5761**†
	Recall@20	0.1046	0.2075	0.2075	0.2168	0.2227	0.1081	0.1442	0.2252	**0.2257**†
	NDCG@20	0.0877	0.1787	0.1787	0.1885	0.1917	0.0853	0.1021	0.1879	**0.1921**†
	PHR@20	0.4392	0.6116	0.6116	0.6326	0.6342	0.4558	0.5378	0.6377	**0.6390**†

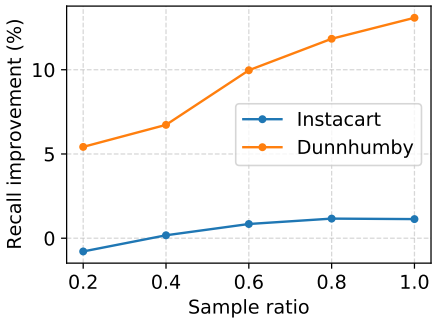


Figure 3.3: The recall improvement of (+T+RF) over (+T) when the training sample ratio changes from 0.2 to 1.

item-specific repetition feature $RepI(i)$ contribute to the accuracy performance of TREx-Rep. Significant improvements over only using the time-decay factor β can be observed on the Dunnhumby dataset when the item-specific repetition feature $RepI(i)$ is also adopted to compute the repetition score $RepS^u(i)$. Note that the improvement of adding $RepI(i)$ to TREx-Rep on the Instacart dataset is relatively small. We conjecture that items in the Instacart dataset are more regular products, that have little difference in repetition feature with each other. Figure 3.3 shows the performance when using different amounts of training samples, the improvement in recall resulting from adding $RepI(i)$ increases when we use more training data since we have more samples for estimating the repetition feature $RepI(i)$.

3.6.3 Beyond-accuracy performance

We conduct experiments to verify whether TREx with the designed models (i.e., TREx$_{Diversity}$ and TREx$_{Fairness}$) could achieve better performance on representative diversity and item fairness metrics. Note that the recommended basket remains fixed for a specific user in existing baselines, resulting in fixed performance regarding both accuracy and beyond-accuracy metrics on each dataset. In contrast, TREx provides the flexibility to adjust the trade-off between accuracy and beyond-accuracy metrics by adjusting the repetition confidence bound v. This allows for a more nuanced control over the recommendation process compared to traditional baselines.

Diversity. The experimental results w.r.t. the accuracy and different diversity metrics (i.e., ILD, Entropy, and DS) are shown in Figure 3.4.[9] We have the following observations: (i) Compared to methods (i.e., TIFUKNN and ReCANet) with the best accuracy, TREx$_{Diversity}$ can achieve better performance in terms of all three diversity metrics while preserving the same level of accuracy on both datasets. (ii) In contrast to other baseline methods (excluding TIFUKNN and ReCANet), TREx$_{Diversity}$ showcases the ability to recommend baskets with enhanced accuracy and diversity simultaneously.

Item fairness. The experimental results regarding the accuracy and five fairness met-

[9] G-TopFreq and Dream exhibit low recall, fairness, and diversity, which prevents them from being visible in Figures 3.4 and 3.5.

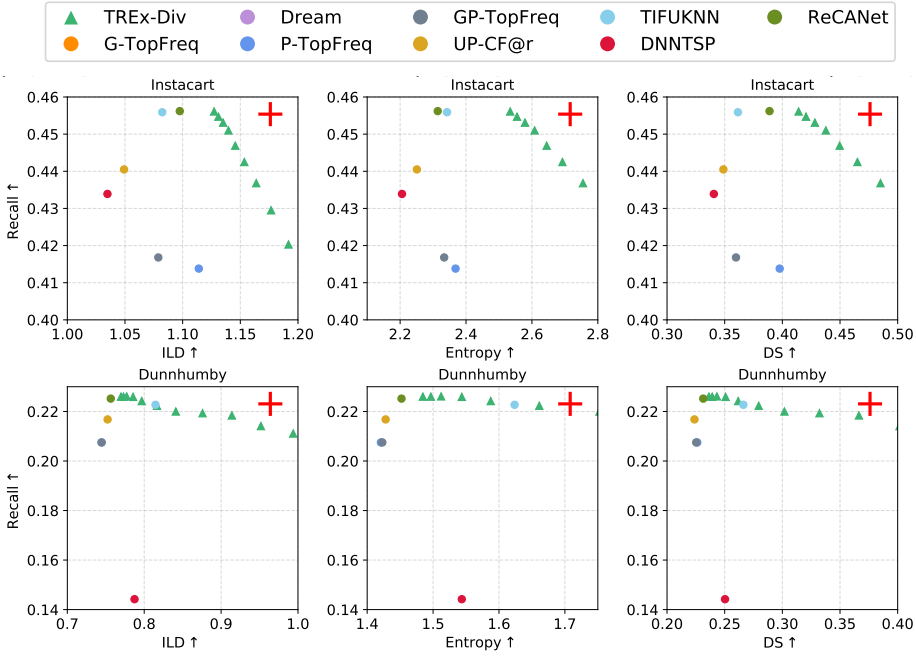


Figure 3.4: Performance of TREx$_{Diversity}$ at different v values, compared with different NBR methods in terms of different diversity metrics. The red $+$ marker indicates the direction with both high accuracy and diversity.

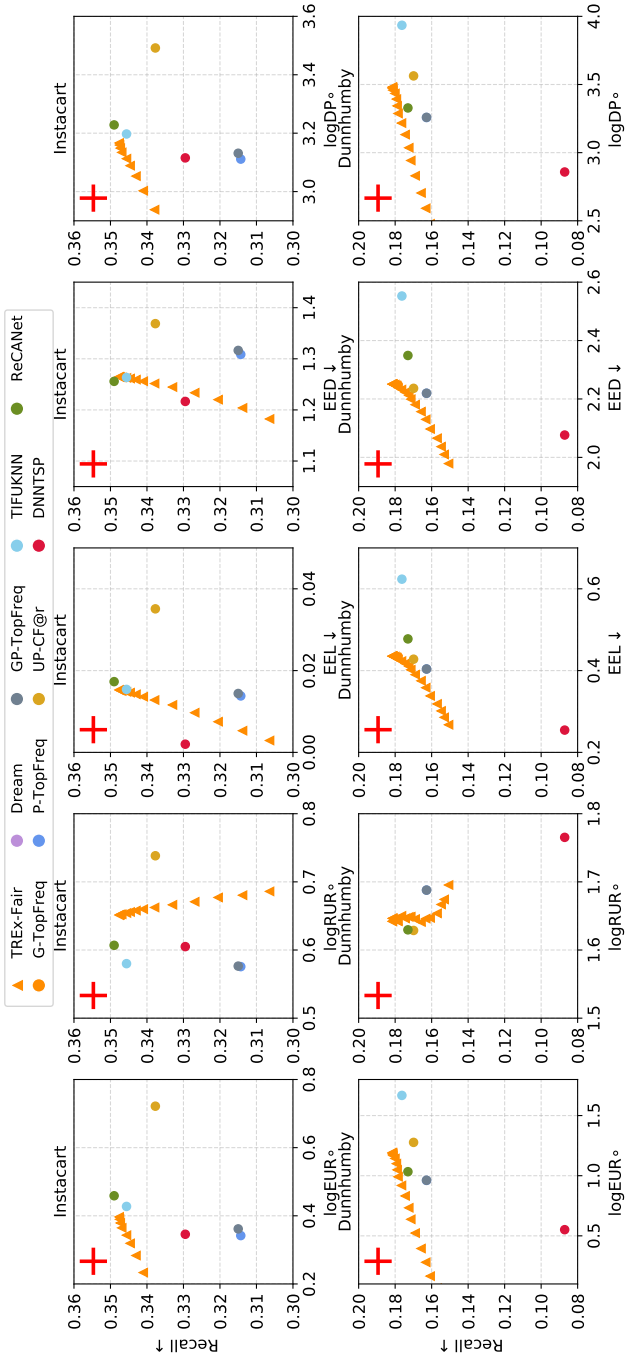


Figure 3.5: Performance of TREx$_{Fairness}$ at different v values, compared with different NBR methods in terms of different fairness metrics. The red $+$ marker indicates the direction with both high accuracy and fairness.

rics (LogRUR, logEUR, logDP, EEL, and EED) are depicted in Figure 3.5. Based on our analysis, we make the following observations: (i) On the Dunnhumby dataset, TREx$_{Fairness}$ demonstrates superior fairness w.r.t. logDP and logEUR while maintaining the same level of accuracy performance as the best-performing baselines (i.e., TIFUKNN and ReCANet). Similarly, on Dunnhumby, TREx$_{Fairness}$ showcases enhanced fairness across four fairness metrics (logDP, logEUR, EEL, and EED) while achieving accuracy performance comparable to the best-performing baselines. (ii) TREx$_{Fairness}$ demonstrates its capability to recommend baskets with improved accuracy and fairness w.r.t. logDP and logEUR concurrently, when compared to complex baselines such as Dream, UP-CF@r, and DNNTSP. (iii) In terms of logRUR, TREx$_{Fairness}$ exhibits inferior performance in fairness while maintaining similar accuracy levels compared to several existing baselines. Moreover, as both accuracy and fairness decrease simultaneously, a win-win and lose-lose scenario is evident rather than a conventional trade-off relationship in this fairness evaluation.

Connections with accuracy. To get a better understanding of the possibility of leveraging the "short-cut" via TREx to improve beyond-accuracy metrics, we conduct an analysis by categorizing these beyond-accuracy metrics into different groups based on their connections with accuracy (see Section 3.4 and Table 3.3).

We can observe that TREx can easily achieve better performance w.r.t. beyond-accuracy metrics that have no connections with accuracy (i.e., ILD, Entropy, DS, and logDP) on two datasets. When beyond-accuracy metrics (e.g., logEUR, EEL, and EED) exhibit weak associations with accuracy, TREx outperforms alternative methods in some instances (4 out of 6). However, in cases where beyond-accuracy metrics are strongly correlated with accuracy (e.g., logRUR), TREx struggles to achieve superior performance. Since only accurate predictions contribute to improvements in logRUR fairness, leveraging the exploration module to optimize such beyond-accuracy metrics is very challenging.

3.6.4 Reflections and discussions

The above results verify our hypothesis and demonstrate the effectiveness of leveraging a "short-cut" strategy to achieve better beyond-accuracy under the current evaluation paradigms.

It is controversial to use this "short-cut" strategy in real-world scenarios when NBR practitioners consider beyond-accuracy metrics. In scenarios where the accuracy of exploration is not important to practitioners and only overall accuracy is of concern, the "short-cut" strategy proves to be a straightforward and efficient means to achieve better performance w.r.t. various beyond-accuracy metrics. TREx must be considered or serve as a baseline before designing more sophisticated methods, such as including multi-objective loss functions [19, 58], integer programming [149], and so on.

However, in some scenarios, it is unreasonable to sacrifice the exploration accuracy despite it being low. Therefore, the existence of the "short-cut" strategy reveals the potential flaws of the existing evaluation paradigms (i.e., using overall metrics to define success). We look into the exploration accuracy [63] of TREx$_{Diversity}$ when it outperforms several existing baselines in terms of both overall accuracy and diversity (i.e., success according to existing evaluation paradigm). Table 3.6 shows the huge de-

crease in the accuracy of exploring items in the recommended basket of TREx$_{Diversity}$, compared to these baselines, since the designed module in TREx$_{Diversity}$ is mainly designed for improving diversity and does not consider accuracy. In this sense, we can not simply claim the superiority of TREx$_{Diversity}$ compared to these baselines just depends on the overall performance.

Note that the fundamental reason for the existence of this "short-cut" is that predicting accurate *explore items* is much more difficult than predicting *repeat items*, and exploration prediction only accounts for a limited user's overall accuracy [60, 62–64]. *Given that exploration prediction contributes only minimally to the overall accuracy of users, it becomes feasible to allocate resources toward optimizing other beyond-accuracy metrics instead of accuracy itself.*

Therefore, beyond using the overall performance to measure accuracy and beyond-accuracy metrics, a fine-grained level evaluation could help to provide a more rigid identification of the success when considering beyond-accuracy metrics.

Table 3.6: Exploration accuracy [63] of TREx$_{Diversity}$ compared with NBR methods that are inferior to it within existing evaluation paradigms.

Dataset	Metric	TIFUKNN	Dream	DNNTSP	TREx-Div
Instacart	$\text{Recall}_{expl}@10$	0.0014	0.0322	0.0014	0.0002
	$\text{PHR}_{expl}@10$	0.0037	0.1431	0.0040	0.0009
	$\text{Recall}_{expl}@20$	0.0077	0.0526	0.0072	0.0008
	$\text{PHR}_{expl}@20$	0.0198	0.2120	0.0217	0.0031
Dunnhumby	$\text{Recall}_{expl}@10$	0.0042	0.0111	0.0017	0.0000
	$\text{PHR}_{expl}@10$	0.0139	0.0521	0.0085	0.0019
	$\text{Recall}_{expl}@20$	0.0069	0.0214	0.0028	0.0016
	$\text{PHR}_{expl}@20$	0.0232	0.1045	0.0115	0.0065

3.7 Conclusion

In this chapter, we have expanded the research objectives of NBR to go beyond sole accuracy to encompass both accuracy and beyond-accuracy metrics to answer RQ2:

> *How can we design a basket recommendation method that jointly targets accuracy and beyond-accuracy objectives, and does the current evaluation paradigm adequately support this goal?*

First, we have recognized a potential "short-cut" strategy to optimize beyond-accuracy metrics while preserving high accuracy levels. To capitalize on and validate the presence of such "short-cuts," we have introduced a plug-and-play framework called *two-step repetition-exploration* (TREx) considering the differences between repetition and exploration tasks. This framework treats repeat items and explore items as distinct entities, employing a straightforward yet highly effective repetition module to uphold accuracy standards. Concurrently, two exploration modules have been devised to target the optimization of beyond-accuracy metrics. We have conducted experiments on two

publicly available datasets w.r.t. eight representative beyond-accuracy metrics, including item fairness (i.e., logEUR, LogRUR, logDP, EEL, and EED) and diversity (i.e., ILD, Entropy, and DS).

Our experimental results demonstrate the effectiveness of our proposed "short-cut" strategy, which can achieve better beyond-accuracy performance w.r.t. several fairness and diversity metrics on different datasets. Additionally, we group beyond-accuracy metrics according to the strength of their connection with accuracy. Our analysis reveals that the stronger the connection with accuracy, the more difficult it becomes to employ a "short-cut" strategy to optimize these beyond-accuracy metrics, favoring the metrics with a stronger connection to avoid such short-cuts.

Implications. As to the broader implications of our work, we have discussed the reasonableness of leveraging the "short-cut" strategy to trade the accuracy of exploration for beyond-accuracy metrics in various scenarios. The presence of this "short-cut" highlights a potential flaw in the definition of success within existing evaluation paradigms, particularly in scenarios where exploration accuracy is important despite being low [122]. A fine-grained level evaluation should be performed in NBR to offer a more precise identification of achieving "better" performance in such a scenario.

Despite the simplicity of the "short-cut" strategy and TREx, our work sheds light on the research direction of considering both accuracy and beyond-accuracy metrics in NBR. Rather than blindly embracing sophisticated methods in NBR, follow-up research should realize the existence of the "short-cut" and potential flaws of existing evaluation paradigms in this research direction.

In the next chapter, we will build on the insights gained in this chapter by specifically mitigating repeat bias while simultaneously optimizing item fairness or diversity for NBR.

3.8 Appendix

To facilitate reproducibility, we release the source code and all hyper-parameters in an online repository: https://github.com/lynEcho/TREX.

4

Repeat-bias-aware Optimization of Beyond-Accuracy Metrics for Next Basket Recommendation

As explained in the previous chapter, in next basket recommendation (NBR) a set of items is recommended to users based on their historical basket sequences. In many domains, the recommended baskets consist of both repeat items and explore items. Some state-of-the-art NBR methods are heavily biased to recommend repeat items so as to maximize utility. The evaluation and optimization of beyond-accuracy objectives for NBR, such as item fairness and diversity, has attracted increasing attention. How can such beyond-accuracy objectives be pursued in the presence of heavy repeat bias? We find that only optimizing diversity or item fairness without considering repeat bias may cause NBR algorithms to recommend more repeat items. This leads us to ask:

RQ3 *How can we design algorithms to jointly improve diversity and item fairness while mitigating repeat bias in next basket recommendation?*

4.1 Introduction

In NBR a recommender system is meant to recommend a set of items at once [66]. In many e-commerce scenarios in which NBR is deployed, users display *repetitive* consumption behavior (e.g., purchasing milk every week, or listening to the same song during workouts) as well as *exploratory* consumption behavior (e.g., buying Christmas gifts) at the same time. Hence, items in a recommended basket can be categorized into repeat items and explore items, depending on whether a user has consumed an item before.

Beyond-accuracy metrics and NBR. Many machine learning techniques have been applied to optimize the accuracy achieved on the NBR task, based on k-nearest neighbors (KNNs) [33, 45, 87], recurrent neural networks (RNNs) [137], or graph neural networks

This chapter was published as Y. Liu, M. Li, M. Aliannejadi, and M. de Rijke. Repeat-bias-aware optimization of beyond-accuracy metrics for next basket recommendation. In *Proceedings of the 47th European Conference on Information Retrieval*, pages 214–229, 2025.

(GNNs) [138]. However, measuring and optimizing beyond-accuracy objectives for NBR remains largely unexplored. Of particular interest are *diversity*, to combat the problem of recommendation homogeneity and generate diversified baskets [58, 108], and *item fairness*, to ensure a fair distribution of the exposure assigned to different groups of items [75]. Optimizing for beyond-accuracy metrics such as diversity and item fairness in the context of NBR is made more complex than in other recommendation scenarios due to the presence of repeat bias.

Repeat bias in NBR. Experiments have shown that repeat items contribute most of the accuracy performance [60, 62–64]. This is because repetition prediction is an easier task, typically with just dozens or hundreds of candidate items and explicit user feedback [65]. However, explore items hold great potential for long-term value in particular and for beyond-accuracy goals in general. E.g., exploring uncertain regions enhances exposure to new and long-tail content, reshaping the overall distribution of the contents, which ultimately improves long-term user experience [105]. Therefore, when curating sets that include both previously seen and new items, it is important to strike the right balance between the repeat and explore categories [4]. Importantly, finding a good balance between repeat and explore items in NBR is made more complex when beyond-accuracy metrics are considered [65]: if most utility is obtained from a relatively small number of items (the repeat items), there is little room to improve beyond-accuracy metrics without sacrificing utility.

Repeat-bias-aware optimization for NBR. For the generic top-k recommendation task, re-ranking is a direct and effective method for improving beyond-accuracy performance, e.g., through greedy algorithms and constrained optimization [149]. The essence of re-ranking is to determine and exploit an effective trade-off between predicted relevance score and beyond-accuracy metrics. However, only optimizing diversity or item fairness may lead to increased repeat bias in some cases according to our experiments, i.e., to an increase in the deviation of the repeat ratio in recommended baskets vs. in ground truth baskets [110]. Even though previous research highlights the problem of repeat bias, no one has tried to optimize to mitigate it. To the best of our knowledge, we are the first to optimize repeat bias jointly with other metrics to seek a balance among (mitigating) repeat bias, accuracy, and beyond-accuracy metrics.

To solve this optimization problem, we propose a model-agnostic repeat-bias-aware diversity optimization (RADiv) algorithm and a repeat-bias-aware item fairness optimization (RAIF) algorithm based on *mixed-integer linear programming* (MILP) for NBR. The repeat ratio (or rather: reducing it) is one of the optimization objectives. The proposed algorithms optimize for predicted relevance, diversity (or item fairness), and repeat ratio simultaneously. We offer several flavors of our MILP-based optimization algorithm, to benefit from the peculiarities of different families of NBR methods. To the best of our knowledge, we are the first to apply a re-ranking algorithm to jointly optimize beyond-accuracy metrics and repeat bias in NBR.

Summarizing, the **main contributions** of the chapter are:

- We propose repeat-bias-aware optimization algorithms (RADiv and RAIF) for NBR, which mitigate repeat bias while improving the diversity and item fairness of recommended baskets.

- We extend these algorithms to multiple NBR paradigms, including ones that merge and optimize items from separate repetition and exploration models.
- We conduct experiments on three retail datasets and find that the proposed RADiv and RAIF algorithms can significantly improve diversity, and item fairness and mitigate repeat bias with an acceptable loss in Recall.

4.2 Related Work

Beyond-accuracy objectives in recommender systems. In recommendation, optimizing only for accuracy measures is limiting and misguided [7]. There is a growing interest in beyond-accuracy metrics, which measure other recommendation qualities [25]. Kaminskas and Bridge [52] study the definitions and metrics of diversity, serendipity, novelty, and coverage. They implement re-ranking strategies for these beyond-accuracy metrics to investigate correlations between the two objectives.

Fairness of recommender systems has sparked much research in the community, considering multiple stakeholders of typical e-commerce platforms. From the user side, user fairness requires that a fair system should provide the same recommendation quality for different user groups [68]. From the item side, the goal is to measure the exposure assigned to each item, or each group, and evaluate this distribution to ensure fair principles, such as statistical parity or equal opportunity [98]. Usually, the exposure of an item in a ranked list is computed based on a user browsing model [14, 83].

Beyond-accuracy objectives in NBR. In NBR, most deep-learning-based models predict the top-k relevant items via a user representation to form a basket. This paradigm leads to the problem of over-homogenization of the recommended baskets [108]. To address this problem, Sun et al. [108] apply an autoregressive item-level decoder to generate items one by one to ensure diversified baskets. Leng et al. [58] employ a deconvolutional network to generate diverse NBR results. Regarding item fairness, Liu et al. [75] reproduce a set of item fairness metrics to evaluate representative NBR methods. Li et al. [65] propose a framework to identify short-cuts in achieving better accuracy and beyond-accuracy performance and advocate fine-grained evaluations in NBR.

Apart from the work listed above, optimizing beyond-accuracy objectives in NBR remains unexplored. Regarding the imbalance between repeat and exploration recommendations, Li et al. [63] first define the repeat ratio of the recommended basket and point out the problem of repeat bias in NBR. Furthermore, Tran et al. [110] formulate repeat bias as the deviation of the repeat ratio of the recommended baskets and ground truth baskets. We contribute by optimizing both diversity and item fairness through re-ranking, while also mitigating repeat bias to improve overall recommendation quality.

Fair/diverse re-ranking. Re-ranking algorithms are usually designed to adjust the ranked results obtained from information access systems considering beyond-accuracy objectives. Commonly used re-ranking algorithms are constrained optimization and maximal marginal relevance (MMR) [11]. For diverse re-ranking, Zhang and Hurley [146] summarize three classic patterns for improving the diversity of recommendation

Table 4.1: Notation used in the chapter.

$u \in U$	Users
$i \in I$	Items
$RepRatio_{rec}$	Repeat ratio of recommendation
$RepRatio_{gt}$	Repeat ratio of ground truth
K	Basket size
N	Number of item candidates
C	Categories of items
W	Decision matrix
W^r	Decision matrix of repeat items
W^e	Decision matrix of explore items
S	Predicted relevance matrix
S^r	Predicted relevance matrix of repeat items
S^e	Predicted relevance matrix of explore items
I_u^{rep}	Repeat item set of user u
I_1	Popular item group
I_2	Unpopular item group
$L_N(u)$	Item candidate list of user u
$L'_K(u)$	New basket of user u after re-ranking
$H(\theta)$	Number of repeat items in combined baskets

lists as constrained optimization problems: (i) maximize the diversity under the constraint of relevance tolerance; (ii) maximize relevance under the constraint of diversity tolerance; and (iii) maximize the weighted sum of relevance and diversity. In terms of fair re-ranking, Biega et al. [9] propose an online optimization approach that uses integer linear programming (ILP) for re-ranking based on accumulated attention and relevance scores while constraining according to a bound on the loss of ranking quality. Singh and Joachims [104] optimize a probabilistic ranking to maximize expected utility under three optional fairness constraints. For fair recommendation, Li et al. [68] set the user fairness metric as a constraint to reduce the recommendation quality gap between the advantaged and disadvantaged groups. CPFair [86] simultaneously optimizes consumer and producer fairness in the objective function for multiple rankings. Compared to previous work, which primarily focuses on fairness or diversity in isolation, we are the first to employ a MILP-based re-ranking algorithm to optimize diversity or item fairness, specifically accounting for repeat bias in NBR task.

4.3 Method

The notation that we use in this chapter is summarized in Table 4.1.

4.3.1 Problem formulation

In NBR, given a user set U and an item set I, for each user $u \in U$, the purchase history is represented as a sequence of baskets $[B_u^1, B_u^2, \ldots, B_u^t]$, where each basket contains a set of items $B_u^t = \{i_1, i_2, \ldots, i_m | i \in I\}$. In NBR the task is to predict the next basket

B_u^{t+1} for each user u. Following the common setting, the size of the recommended basket is fixed as K. Generally, NBR methods generate an item ranking $L(u)$ for each user based on the predicted relevance score S_{ui}, and then select the top-K relevant items to form the next basket, i.e., $B_u^{t+1} = L_K(u)$.

4.3.2 Model overview

In the proposed re-ranking procedure, we select top-N relevant items $L_N(u)$ for each user as candidates. The relevance scores of candidate items of all users can be represented as $S = [S_{ui}]_{|U|\times N}$. A new basket $L'_K(u)$ is selected from the candidate list $L_N(u)$ taking into account the predicted relevance, as well as the diversity, item fairness, and repeat bias of the basket. $L'_K(u) \subset L_N(u), N > K$. The re-ranking procedure can be formalized as a mixed-integer linear programming (MILP) problem, which is a mathematical optimization problem that can be solved by heuristic algorithms and optimization solvers. In our MILP formulation, the objective function is designed to maximize relevance, diversity (or item fairness) and reduce repeat bias simultaneously. W is a binary decision matrix, $W = [W_{ui}]_{|U|\times N}$. The element $W_{ui} = 1$ indicates recommending item i to user u, and 0 otherwise. The proposed RADiv and RAIF algorithms described below are specifically adapted for different families of NBR methods: unified methods and combined methods.

4.3.3 Optimization objectives

In this section, we introduce the definition of three optimization objectives that we aim to achieve.

Diversity. Diverse recommendation aims to recommend items of various and different categories to users. In this chapter, we use the diversity score (DS) [71] as the diversity objective, computed by dividing the number of recommended categories by the basket size K given the decision matrix W. Diversity objective $\mathrm{DS}(C, W)$ is the sum of DS among all users:

$$\mathrm{DS}(C, W) = \sum_{u\in U} \frac{(\#\text{recommended categories}|W)}{K}. \tag{4.1}$$

Item fairness. We evaluate the item fairness between popular item group I_1 and unpopular group I_2. The recommended items will receive exposure related to the position. Demographic parity (DP) [104] defines fairness of exposure as equal average exposure between the two groups. The average exposure of a group I_k is computed as:

$$Exposure(I_k \mid W) = \frac{1}{|I_k|} \sum_{i\in I_k} Exposure(i \mid W). \tag{4.2}$$

Inspired by the construction in [86], the item fairness objective is designed as the difference of average exposure between group I_1 and I_2. The closer the IF value is to zero, the fairer a ranking is:

$$\mathrm{IF}(I_1, I_2, W) = Exposure(I_1 \mid W) - Exposure(I_2 \mid W). \tag{4.3}$$

Repeat bias. Some NBR methods are either biased to recommending too many repeat items or explore items compared with the ground truth baskets. We aim to mitigate

the repeat bias of recommended baskets. Since the repeat ratio of ground truth baskets $RepRatio_{gt}$ is unknown, we directly choose the repeat ratio of recommended baskets $RepRatio_{rec}$ as the optimization objective. Each user u has a repeat item set I_u^{rep}, which contains all the items the user has bought before. Given the decision matrix W, the repeat ratio objective can be expressed as:

$$RepRatio_{rec}(I_u^{rep}, W) = \sum_{u \in U} \frac{(|L'_K(u) \cap I_u^{rep}| \, | W)}{K} \, . \tag{4.4}$$

4.3.4 Repeat-bias-aware optimization for unified NBR methods

In this section, we post-process the recommendation lists of users generated by unified NBR methods, where a unified model generates the relevance scores of all items. We select top-N relevant items of each user as candidates. The relevance scores of these item candidates can be represented as $S = [S_{ui}]_{|U| \times N}$. We apply a MILP model to re-rank the top N candidates for each user. The repeat-bias-aware diversity optimization (RADiv) algorithm simultaneously optimizes diversity and repeat ratio as shown in Eq. 4.5:

$$\begin{aligned} &\max \frac{1}{K} \sum_{u \in U} \sum_{i=1}^{N} S_{ui} W_{ui} + \epsilon_1 \, \mathrm{DS}(C, W) - \lambda RepRatio_{rec}(I_u^{rep}, W) \\ &\text{such that } \sum_{i=1}^{N} W_{ui} = K, W_{ui} \in \{0, 1\}. \end{aligned} \tag{4.5}$$

We design the objective function to maximize the sum of relevance scores, the diversity score $\mathrm{DS}(C, W)$, while minimizing the $RepRatio_{rec}(I_u^{rep}, W)$. Here, ϵ_1 and λ are the weighting parameters of the diversity term and repeat ratio term, respectively. W is a binary decision matrix determining whether to recommend item i to user u. The constraint indicates that the algorithm ultimately recommends K items to each user, which is the basket size. It is worth noting that the minus sign in front of $RepRatio_{rec}$ is designed for repeat-biased NBR methods with quite high $RepRatio_{rec}$. The minus implies that the algorithm aims to reduce the $RepRatio_{rec}$ to be closer to $RepRatio_{gt}$. In contrast, for explore-biased methods with quite low $RepRatio_{rec}$ values, this term should be set to a positive sign: $+\lambda RepRatio_{rec}(I_u^{rep}, W)$. The algorithm will increase the $RepRatio_{rec}$ so as to approach $RepRatio_{gt}$.

Similarly, the repeat-bias-aware item fairness optimization (RAIF) algorithm simultaneously optimizes item fairness and repeat ratio by Eq. 4.6.

$$\begin{aligned} &\max \sum_{u \in U} \sum_{i=1}^{N} S_{ui} W_{ui} - \alpha_1 \, \mathrm{IF}(I_1, I_2, W) - \lambda RepRatio_{rec}(I_u^{rep}, W) \\ &\text{such that } \sum_{i=1}^{N} W_{ui} = K, W_{ui} \in \{0, 1\}. \end{aligned} \tag{4.6}$$

The objective function is designed to maximize the combination of relevance scores, item fairness $\mathrm{IF}(I_1, I_2, W)$, and minimize $RepRatio_{rec}(I_u^{rep}, W)$. Here, α_1 and λ are the weighting parameters of item fairness term and repeat ratio term, respectively.

4.3.5 Repeat-bias-aware optimization for combined NBR methods

In the previous setting, the optimization algorithm is designed for unified NBR methods, where the relevance scores of repeat items and explore items are generated by the same model and are comparable. However, there are another NBR paradigms, such as two-step repetition-exploration (TREx) [65] (see Chapter 3), where the repeat item list and explore item list are obtained from different models. This paradigm allows one to combine the strongest repetition model and the strongest exploration model flexibly, and adjust the proportion of these two parts. When we optimize diversity (or item fairness) and repeat bias of combined NBR methods, the challenge lies in the fact that the predicted relevance scores are not comparable between repeat items and explore items.

Inspired by the TREx framework [65], we use a threshold θ to filter the repeat item candidates and count the number of repeat items with a relevance score larger than θ. The numbers of repeat items for each user are saved as vector $H(\theta)_{\text{aux}}$. The final repeat position number for each user is $H(\theta) = \min(H(\theta)_{\text{aux}}, K)$. Here, θ indirectly controls $RepRatio_{rec}(\theta)$. The higher the θ becomes, the lower the number of repeat slots $H(\theta)$ and $RepRatio_{rec}(\theta)$ become. $H(\theta)$ is an important variable to avoid the comparison between repeat and explore relevance scores. In this way, repeat items are compared internally and $H(\theta)$ of them are finally selected. Explore items compete internally and $K - H(\theta)$ of them are chosen.

The RADiv algorithm is adapted for combined NBR methods as shown in Eq. 4.7. $S^r = [S^r_{ui}]_{|U|\times N}$ and $S^e = [S^e_{ui}]_{|U|\times N}$ are predicted relevance score matrices of repeat items and explore items from different recommender systems, respectively. $W^r = [W^r_{ui}]_{|U|\times N}$, $W^e = [W^e_{ui}]_{|U|\times N}$ are corresponding binary decision matrices to determine the selection of repeat items and explore items. ϵ_2 is the weighting parameter of the diversity term. The constraints indicate the repeat slots for each user are $H(\theta)$, while the explore slots for each user are $K - H(\theta)$. Similarly, RAIF is adjusted to optimize item fairness $\text{IF}(I_1, I_2, W^r, W^e)$ and $RepRatio_{rec}(\theta)$ in Eq. 4.8. α_2 is the weighting parameter of item fairness term. Algorithm 2 summarizes our algorithms for combined NBR methods.

$$\begin{aligned} &\max \frac{1}{K} \sum_{u\in U} \sum_{i=1}^{N} (S^r_{ui} W^r_{ui} + S^e_{ui} W^e_{ui}) + \epsilon_2 \,\text{DS}(C, W^r, W^e) \\ &\quad \text{such that } \sum_{i=1}^{N} W^r_{ui} = H(\theta), \sum_{i=1}^{N} W^e_{ui} = K - H(\theta), W^r_{ui}, W^e_{ui} \in \{0,1\}; \end{aligned} \tag{4.7}$$

$$\begin{aligned} &\max \sum_{u\in U} \sum_{i=1}^{N} (S^r_{ui} W^r_{ui} + S^e_{ui} W^e_{ui}) - \alpha_2 \,\text{IF}(I_1, I_2, W^r, W^e) \\ &\quad \text{such that } \sum_{i=1}^{N} W^r_{ui} = H(\theta), \sum_{i=1}^{N} W^e_{ui} = K - H(\theta), W^r_{ui}, W^e_{ui} \in \{0,1\}. \end{aligned} \tag{4.8}$$

Algorithm 2: RADiv and RAIF for Combined NBR Methods

Input: Basket size K, Categories C, Number of item candidates N, Item group I_1, I_2, Parameters $\epsilon_2, \alpha_2, \theta$
Output: Recommendation matrix W^r, W^e
1 $S^r, S^e \leftarrow$ The top-N repeat and explore relevance scores of users.
2 Compare θ with S^r_{ui}, and obtain $H(\theta) = \min(H(\theta)_{\text{aux}}, K)$.
3 Solve the optimization problem following Eq. 4.7 (or 4.8).
4 **Return** W^r, W^e

4.4 Experimental Setup

NBR methods. We select and investigate the following five representative NBR methods:

- **UP-CF@r**, which combines recency-aware user-wise popularity and collaborative filtering while considering the recent shopping behavior [33].
- **TIFUKNN**, which models the temporal dynamics of users' past baskets by using a KNN-based approach based on the personal item frequency (PIF) [45].
- **Dream**, which forms basket representations using a pooling strategy and models sequential user behavior through an RNN [137].
- **DNNTSP** uses GNN and self-attention mechanisms to encode item-to-item relations across baskets and capture temporal dependencies [138].
- **TREx**, which has a repetition module considering item repurchase features and users' interests, and an exploration module targeted for beyond-accuracy metrics. Then, repeat items and explore items are combined to form the final basket [65].

We exclude NBR methods that only focus on the repetition recommendation (i.e., P-TopFreq, ReCANet [3], and NBRR [54]) or exploration recommendation (i.e., NNBR [61]), as we aim to investigate and balance repetition *and* exploration in NBR.

Datasets. Following previous NBR studies [61, 63, 65, 75], we select three publicly available grocery shopping datasets: TaFeng [109], Dunnhumby [29], and Instacart [48]. They exhibit different characteristics in terms of repetition and exploration, which are critical for verifying the effectiveness of our proposed repeat-bias-aware optimization algorithms.

For each dataset, we remove users with fewer than three baskets and items purchased fewer than five times, as done in [3]. Due to the large size of the Instacart dataset, memory limitations occurred during the calculation of some methods. Following [87], we also randomly sampled 20,000 users from Instacart before applying any filtering. Table 4.2 provides the statistics of the three datasets after preprocessing. The average $RepRatio_{gt}$ refers to the average proportion of repeat items in the ground truth baskets, as defined in [63].

Table 4.2: Statistics of the datasets after preprocessing.

Dataset	#Users	#Items	#Baskets	Avg. #baskets/user	Avg. #items/basket	Avg. $RepRatio_{gt}$
Instacart	19,210	29,398	305,582	15.91	10.06	0.60
Dunnhumby	2,482	37,161	107,152	43.17	10.07	0.43
TaFeng	10,182	15,023	82,387	8.09	6.14	0.21

We split each dataset following the approach in [3, 33, 87]. The training set includes all baskets from each user except the last basket. For users with more than 50 baskets in the training data, we limit the training set to their most recent 50 baskets [63]. The last baskets of all users are then randomly split equally into a validation set (50%) and a test set (50%).

Evaluation metrics. We use the following widely used metrics in our experiments to measure how the proposed algorithms balance multiple objectives. In terms of accuracy, Recall measures the system's ability to retrieve items that users will purchase in their next baskets. For item fairness, demographic parity (DP) measures the ratio of the average exposure of the popular group to the average exposure of the unpopular group. Following [98], we use logDP to deal with the empty-group case. The closer the logDP value is to zero, the fairer the recommendation is. The diversity score (DS) measures the number of categories within the recommended basket divided by basket size. Repeat bias is computed as: RepBias = $RepRatio_{rec} - RepRatio_{gt}$ [110]. Following [86], we use a comprehensive metric to evaluate the overall performance of item fairness and RepBias: mFR = $\omega|\text{logDP}| + (1-\omega)|\text{RepBias}|$. Similarly, the overall performance of diversity and RepBias is evaluated as mDR = $\omega\text{DS} - (1-\omega)|\text{RepBias}|$.

Implementation details. In our experiments, we follow [45, 65] and set the basket size $K = 20$. The number of item candidates $N = 100$. We compute the popularity of all items and select the top 20% as popular I_1 and the rest belong to the unpopular group I_2. We apply grid search to select hyperparameters on the validation set. We select α_1 and α_2 in [0, 0.001, 0.01, 0.1, 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200]. ϵ_1 and select ϵ_2 from [0, 0.001, 0.01, 0.02, 0.04, 0.06, 0.08, 0.1, 0.12, 0.14, 0.16, 0.18, 0.2]. We choose λ in [0, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]. θ candidates are selected based on S^r.

We select the optimal hyperparameter combination according to these two conditions: (i) satisfy a 10% Recall drop tolerance; and (ii) maximize mDR or minimize mFR, where $\omega = 0.5$ following [86]. We use Gurobi,[1] which is an industrial optimization solver capable of delivering practical and effective feasible solutions.

4.5 Experimental Results

We design experiments to address the following three sub-research questions: **RQ3.1** How does optimizing only for diversity or item fairness affect the repeat bias? **RQ3.2** What is the performance of the proposed RADiv and RAIF algorithms on NBR methods?

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

RQ3.3 How does RADiv and RAIF algorithms strike a balance between utility and beyond-accuracy metrics?

We start by studying how optimizing for beyond-accuracy metrics, without taking the repeat bias into account, would affect the baskets in terms of repeat-explore items, answering **RQ3.1**. We apply a re-ranking algorithm without considering $RepRatio_{rec}$ to optimize the recommended baskets obtained from different NBR methods on Instacart, Dunnhumby, and TaFeng. For diverse re-ranking, the optimization objective is formulated as $\max \frac{1}{K} \sum_{u \in U} \sum_{i=1}^{N} S_{ui} W_{ui} + \epsilon \, \mathrm{DS}(C, W)$. In terms of fair re-ranking, the objective function is $\max \sum_{u \in U} \sum_{i=1}^{N} S_{ui} W_{ui} - \alpha \, \mathrm{IF}(I_1, I_2, W)$. The parameters ϵ and α are used to adjust the weight of the diversity and item fairness terms, respectively. Figure 4.1 shows the performance of diversity and item fairness optimization on Dunnhumby.[2]

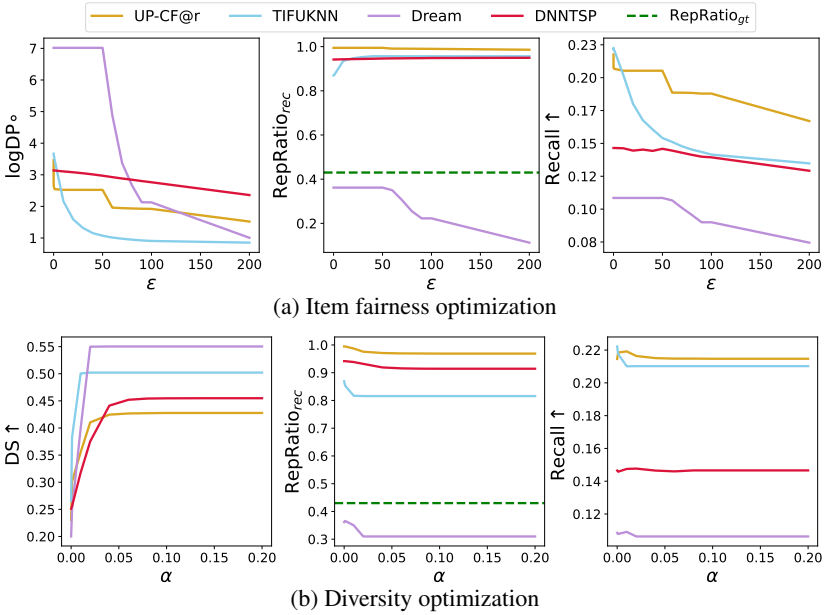
We make the following observations: (i) Diversity and item fairness optimizations significantly improve the NBR methods in terms of logDP and DS. For item fairness optimization, as the parameter ϵ increases ($0 \rightarrow 200$), all NBR methods become fairer. Also, the DS of all NBR methods increase as α increases ($0 \rightarrow 0.2$). (ii) In most cases, Recall exhibits a declining trend as DS and logDP become better. In diversity optimization, the Recall of UP-CF@r, Dream, and DNNTSP improve slightly when $\alpha = 0.01$. (iii) In terms of repeat bias, which is measured as the gap between $RepRatio_{rec}$ and $RepRatio_{gt}$ (green dashed line in Figure 4.1). We see that the repeat bias is likely to either intensify or diminish, showing no clear trend. In item fairness optimization, the repeat bias of TIFUKNN and Dream grows as ϵ increases, showing that it is important to consider repeat bias while optimizing beyond-accuracy metrics.

To answer **RQ3.2**, we apply the RADiv and RAIF algorithms to five NBR methods. Of these, TIFUKNN, UP-CF@r, DNNTSP, and Dream are unified repeat-explore recommendation methods while TREx is a combined method. Since TREx does not provide predicted relevance scores for explore items, here we use the explore item list obtained from UP-CF@r to combine with repeat items of TREx. Table 4.3 reports the results on Instacart.[3] We also conduct an ablation study, i.e., optimizing only for diversity, item fairness, and repeat bias, respectively.

We arrive at the following conclusions: (i) The proposed RADiv algorithm (RD) achieves a significant improvement in DS, RepBias, and the comprehensive metric mDR in comparison to the original (Ori.) results on all NBR methods. Surprisingly, the Recall of Dream increases after RADiv optimization ($0.0977 \rightarrow 0.1468$). (ii) The RAIF algorithm (RF) achieves better logDP, less RepBias, and better overall metric mFR than the original (Ori.) results, proving the effectiveness of RAIF algorithm. (iii) RADiv and RAIF algorithms can effectively adjust the $RepRatio_{rec}$ of methods to approach $RepRatio_{gt}$. Take Instacart as an example, RADiv and RAIF algorithms decrease the $RepRatio_{rec}$ of TIFUKNN, UP-CF@r, DNNTSP, and increase the $RepRatio_{rec}$ of Dream to make them closer to ground truth 0.6. (iv) In our ablation study, we find that only optimizing $RepRatio_{rec}$ (R) will decrease the diversity and item fairness of TIFUKNN, UC-CF@r, DNNTSP and TREx. This means that it is necessary to optimize

[2] We observe a similar trend on the Instacart and TaFeng datasets, and report their results in the Appendix of this chapter.

[3] Similar patterns are observed on the Dunnhumby and TaFeng datasets. We report their results in the Appendix of this chapter.



(a) Item fairness optimization

(b) Diversity optimization

Figure 4.1: Performance of item fairness and diversity optimization without considering $RepRatio_{rec}$ on different NBR methods. logDP∘ means the closer logDP value is to zero, the fairer it is.

both $RepRatio_{rec}$ and diversity, both $RepRatio_{rec}$ and item fairness, as implemented by the proposed RADiv and RAIF algorithms.

To answer **RQ3.3**, we take UP-CF@r as an example to illustrate the trade-off between Recall and other metrics on Instacart. For the two parameters ϵ_1 and λ in the RADiv optimization, we change one parameter each time with the other one fixed. Similarly, we perform the same operation for α_1 and λ in RAIF optimization. The analysis is shown in Figure 4.2.

We make the following observations: (i) In Figure 4.2a, we first fix λ and change ϵ_1 from 0 to 0.2. We see that DS continuously increases at the cost of Recall. The red star indicates that the algorithm chooses $\epsilon_1 = 0.2$ as the optimal solution to achieve a balance between diversity and Recall. When we fix ϵ_1 and change λ from 0 to 1, $RepRatio_{rec}$ and Recall decline consistently. Since our algorithm optimizes $RepRatio_{rec}$ to be approaching $RepRatio_{gt}$, it selects $\lambda = 0.01$ as the optimal trade-off between Recall and repeat bias. This process reflects the balanced strategy of RADiv to maximize diversity and mitigate repeat bias within the tolerance of utility loss. (ii) In Figure 4.2b, we fix λ and adjust α_1 from 0 to 200. logDP decreases indicating unpopular items are assigned with more exposure. The recommendation becomes fairer at the cost of a Recall drop. The algorithm chooses $\alpha_1 = 200$ to achieve a balance between logDP and Recall. When we change λ from 0 to 1 with a fixed α_1, $RepRatio_{rec}$ and Recall show a consistent downward trend. The algorithm selects $\lambda = 0.9$ to balance repeat bias and Recall. Considering the inverse relationship between Recall and logDP, repeat bias, the

Table 4.3: The performance of RADiv and RAIF algorithms on Instacart ($RepRatio_{gt} = 0.6$). *Ori.* indicates the original baskets obtained from each method. *RD* indicates RADiv. *RF* refers to RAIF. In ablation study, *D*, *F*, *R* indicate optimizing only for diversity, item fairness, and repeat bias, respectively. RepR indicates $RepRatio_{rec}$ (best close to $RepRatio_{gt}$). RepBias and logDP (best close to zero).

Meth.	Diversity optimization						Item fairness optimization					
	Type	Recall↑	DS↑	RepR	RepBias	mDR↑	Type	Recall↑	logDP	RepR	RepBias	mFR↓
TIFUKNN	Ori.	**0.4559**	0.3615	0.9248	0.3248	0.0184	Ori.	**0.4559**	3.1252	0.9248	0.3248	1.7250
	D	0.4259	0.5897	0.8984	0.2984	0.1457	F	0.4269	2.3510	0.9252	0.3252	1.3381
	R	0.4537	0.3587	0.9100	0.3100	0.0244	R	0.4232	3.2760	**0.7939**	**0.1939**	1.7350
	RD	0.4245	**0.5898**	**0.8874**	**0.2874**	**0.1512**	RF	0.4098	**2.3271**	0.8718	0.2718	**1.2995**
UP-CF@r	Ori.	**0.4405**	0.3489	0.8905	0.2905	0.0292	Ori.	**0.4405**	3.3966	0.8905	0.2905	1.8436
	D	0.3983	**0.6375**	0.7896	0.1896	0.2239	F	0.4282	2.4860	0.8858	0.2858	1.3859
	R	0.4353	0.3424	0.8812	0.2812	0.0306	R	0.4373	3.3997	0.8791	0.2791	1.8394
	RD	0.3968	**0.6375**	**0.7837**	**0.1837**	**0.2269**	RF	0.4052	**2.1810**	**0.8019**	**0.2019**	**1.1915**
DNNTSP	Ori.	**0.4347**	0.3402	0.9133	0.3133	0.0135	Ori.	**0.4347**	3.2573	0.9133	0.3133	1.7853
	D	0.4046	0.6009	0.8660	0.2660	0.1675	F	0.4332	**2.9780**	0.9152	0.3152	1.6466
	R	0.4337	0.3388	0.9080	0.3080	0.0154	R	0.4256	3.3204	**0.8640**	**0.2640**	1.7922
	RD	0.4059	**0.6013**	**0.8623**	**0.2623**	**0.1695**	RF	0.4277	2.9901	0.8837	0.2837	**1.6369**
Dream	Ori.	0.0977	0.1000	0.1923	-0.4077	-0.1539	Ori.	0.0977	7.3111	0.1923	-0.4077	3.8594
	D	0.0723	**0.7000**	0.1267	-0.4733	0.1133	F	0.0709	**2.1018**	0.1226	-0.4774	**1.2896**
	R	**0.1540**	0.1243	**0.4202**	**-0.1798**	-0.0277	R	**0.1288**	7.3111	**0.3108**	**-0.2892**	3.8002
	RD	0.1468	0.5654	**0.4202**	**-0.1798**	**0.1928**	RF	0.0939	2.3893	0.2094	-0.3906	1.3900
TREx	Ori.	**0.4595**	0.3533	0.9265	0.3265	0.0134	Ori.	**0.4595**	3.2182	0.9265	0.3265	1.7724
	D	0.4455	0.5296	0.9265	0.3265	0.1016	F	0.4569	2.9052	0.9265	0.3265	1.6159
	R	0.4368	0.3280	**0.7973**	**0.1973**	0.0654	R	0.4277	3.4281	**0.7487**	**0.1487**	1.7884
	RD	0.4195	**0.5874**	**0.7973**	**0.1973**	**0.1951**	RF	0.4226	**2.6528**	**0.7487**	**0.1487**	**1.4008**

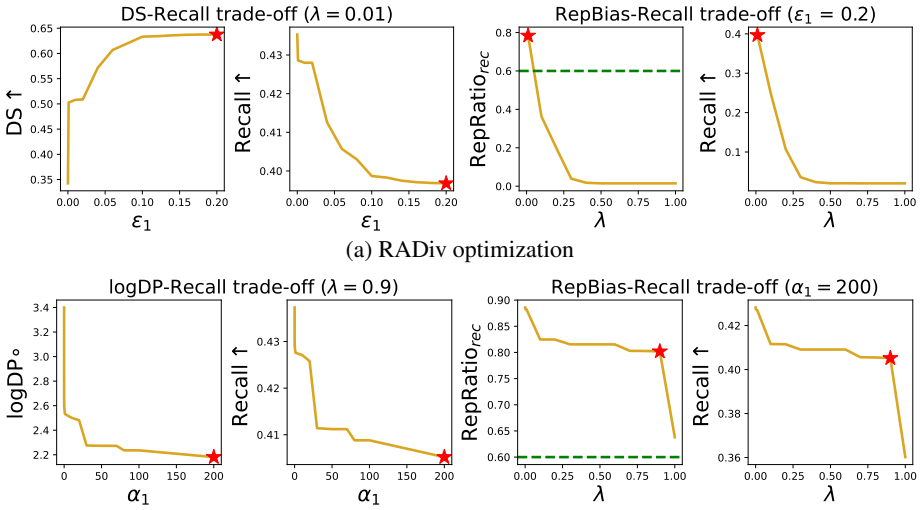
RAIF algorithm decides to sacrifice a little utility in exchange for a greater return on item fairness and repeat bias.

4.6 Conclusion

In this chapter, we focus specifically on measuring and mitigating repeat bias in NBR, while also considering beyond-accuracy evaluation criteria, to address RQ3:

> *How can we design algorithms to jointly improve diversity and item fairness while mitigating repeat bias in next basket recommendation?*

Specifically, we have proposed repeat-bias-aware optimization algorithms for improving diversity and item fairness while mitigating repeat bias in NBR. Our RADiv and RAIF algorithms re-rank the preliminary recommended results obtained from various NBR baselines according to different beyond-accuracy objectives and seek a balance between repeat items and explore items simultaneously while optimizing for beyond-accuracy metrics. We have extended our approach to multiple NBR paradigms [65], in particular, one that fuses repeat item lists and explores item lists from separate models, allowing us to select tailored models for each list. We have conducted experiments



(a) RADiv optimization

(b) RAIF optimization

Figure 4.2: Trade-off between Recall and diversity, item fairness, repeat bias. Take UP-CF@r as an example on the Instacart dataset. The red star indicates the optimal solution chosen by the proposed algorithm. The green dashed line is value of $RepRatio_{gt}$.

on three real-world retail datasets. We find that only optimizing for diversity or item fairness will increase repeat bias in many cases, which can reduce user satisfaction. The proposed RADiv and RAIF algorithms can effectively optimize diversity and fairness while mitigating the repeat bias issue under an acceptable utility loss. Finally, we have investigated trade-offs between Recall and other beyond-accuracy metrics, including diversity, item fairness, and repeat bias.

Implications. Based on our experiments, we find that it is critical to evaluate the utility of NBR methods and measure their repeat bias at the same time. In many cases, a recommendation model can achieve the best performance in terms of Recall by recommending only repeat items (i.e., highest repeat bias), but this reduces the likelihood of novelty and serendipity in recommendations, leading to filter bubble and echo chamber issues [36, 39]. Re-ranking may be an effective solution to this problem, as we have shown that it can strike a balance between accuracy and beyond-accuracy metrics, while taking into account the repeat bias in the re-ranking process. However, in extreme cases where the original ranking does not include many explore items, re-ranking cannot really be effective. Therefore, we suggest to consider accuracy and beyond-accuracy objectives in the original ranking.

In addition, the definition and quantification of repeat bias remain an open question. In this work, we directly use the repeat ratio of the ground truth baskets as a proxy; however, it is worth asking whether other reasonable personalized proportions exist. Since the repeat ratio is an unavoidable factor when evaluating both accuracy and beyond-accuracy metrics, a simpler yet cleaner alternative is to split NBR into two separate tasks: exploratory basket recommendation and repetitive basket recommendation. In this way, we eliminate repeat bias as a confounding factor. In Chapter 5, we focus

exclusively on exploratory basket recommendation.

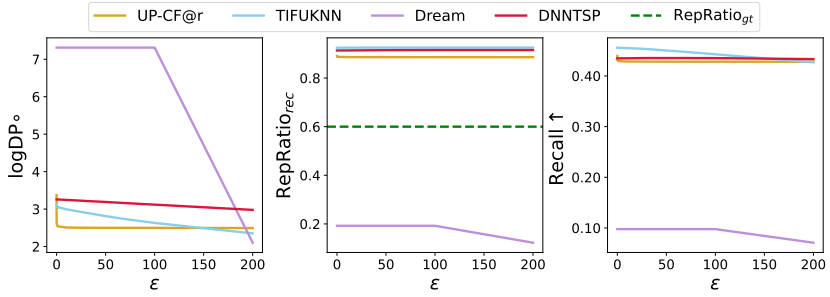
4.7 Appendix

4.7.1 Reproducibility

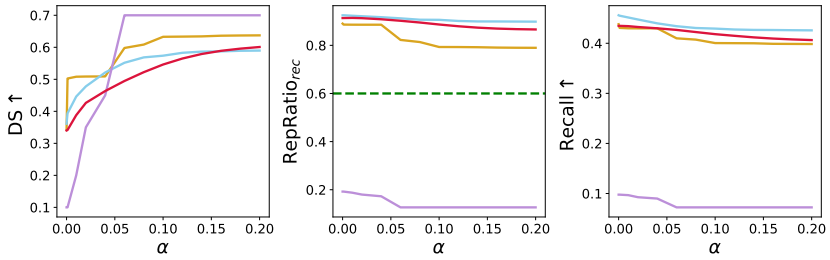
We release our code and hyperparameters at https://github.com/lynEcho/Repbias_NBR.

4.7.2 Additional results

We include additional plots Figure 4.3 and 4.4 to supplement Figure 4.1. We include additional tables Table 4.4 and 4.5 to supplement Table 4.3.

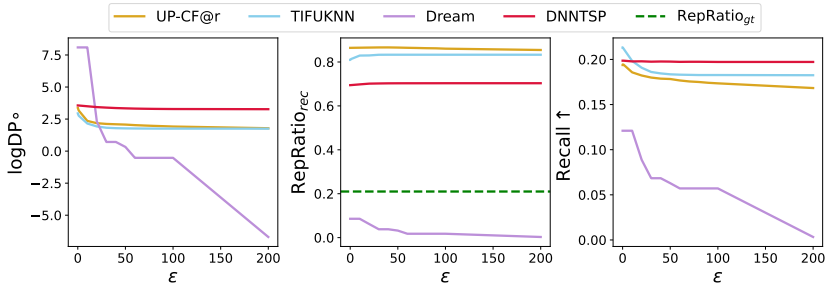


(a) Item fairness optimization

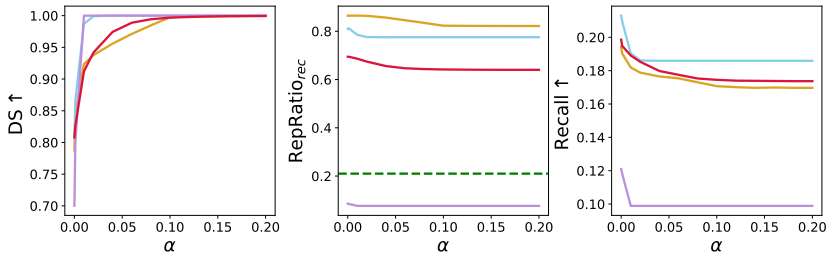


(b) Diversity optimization

Figure 4.3: Performance of item fairness and diversity optimization without considering $RepRatio_{rec}$ on Instacart.



(a) Item fairness optimization



(b) Diversity optimization

Figure 4.4: Performance of item fairness and diversity optimization without considering $RepRatio_{rec}$ on TaFeng.

Table 4.4: The performance of RADiv and RAIF algorithms on Dunnhumby ($RepRatio_{gt} = 0.43$). *Ori.* indicates the original baskets obtained from each method. *RD* indicates RADiv. *RF* refers to RAIF. In ablation study, *D*, *F*, *R* indicate optimizing only for diversity, item fairness, and repeat bias, respectively. RepR indicates $RepRatio_{rec}$ (best close to $RepRatio_{gt}$). RepBias and logDP (best close to zero).

Meth.	Diversity optimization						Item fairness optimization					
	Type	Recall↑	DS↑	RepR	RepBias	mDR↑	Type	Recall↑	logDP	RepR	RepBias	mFR↓
TIFUKNN	N	**0.2227**	0.2662	0.8691	0.4391	-0.0865	N	**0.2227**	3.6680	0.8691	0.4391	2.0536
	D	0.2101	**0.5024**	0.8157	0.3857	0.0583	F	0.2205	**3.4719**	0.8740	0.4440	**1.9580**
	R	0.2088	0.2755	0.7131	0.2873	-0.0059	R	0.2213	3.6793	**0.8632**	**0.4332**	2.0563
	RD	0.2040	**0.5024**	**0.7045**	**0.2745**	**0.1139**	RF	0.2194	3.481	0.8686	0.4386	1.9598
UP-CF@r	N	**0.2168**	0.2238	0.9945	0.5645	-0.1704	N	**0.2168**	3.5031	0.9945	0.5645	2.0338
	D	0.2153	0.4276	0.9692	0.5392	-0.0558	F	0.2049	2.5224	0.9945	0.5645	1.5435
	R	0.2104	0.2321	0.9411	0.5111	-0.1395	R	0.2132	3.4776	**0.9187**	**0.4887**	1.9832
	RD	0.2123	**0.4280**	**0.9358**	**0.5058**	**-0.0389**	RF	0.2037	**2.4892**	0.9747	0.5447	**1.5170**
DNNTSP	N	0.1465	0.2512	0.9416	0.5116	-0.1302	N	**0.1465**	3.1376	0.9416	0.5116	1.8246
	D	**0.1466**	**0.4548**	0.9143	0.4843	-0.0148	F	0.1292	**2.3589**	0.9494	0.5194	1.4392
	R	0.1434	0.2519	0.9160	0.4860	-0.1170	R	0.1411	3.1769	**0.8639**	**0.4339**	1.8054
	RD	0.1453	**0.4548**	**0.8981**	**0.4681**	**-0.0067**	RF	0.1244	2.3727	0.9055	0.4755	**1.4241**
Dream	N	0.1085	0.2000	0.3618	-0.0682	0.0659	N	0.1085	7.0173	0.3618	-0.0682	3.5428
	D	0.1063	**0.5503**	0.3098	-0.1202	0.2151	F	0.0745	**1.0075**	0.1127	-0.3173	**0.6624**
	R	**0.1275**	0.2157	0.6321	0.2021	0.0068	R	**0.1341**	6.9063	0.7440	0.3140	3.6102
	RD	0.1133	**0.5503**	**0.4505**	**0.0205**	**0.2649**	RF	0.1006	1.9855	**0.3830**	**-0.0470**	1.0163
TREx	N	**0.2257**	0.2325	0.9953	0.5653	-0.1664	N	0.2257	3.3947	0.9953	0.5653	1.9800
	D	0.2209	0.4079	0.9953	0.5653	-0.0787	F	**0.2258**	3.3116	0.9953	0.5653	1.9385
	R	0.2033	0.2224	**0.6920**	**0.2620**	-0.0198	R	0.2006	3.7244	**0.6443**	**0.2143**	1.9694
	RD	0.1954	**0.4948**	**0.6920**	**0.2620**	**0.1164**	RF	0.1988	**3.1603**	**0.6443**	**0.2143**	**1.6873**

Table 4.5: The performance of RADiv and RAIF algorithms on TaFeng ($RepRatio_{gt} = 0.21$). *Ori.* indicates the original baskets obtained from each method. *RD* indicates RADiv. *RF* refers to RAIF. In ablation study, *D*, *F*, *R* indicate optimizing only for diversity, item fairness, and repeat bias, respectively. RepR indicates $RepRatio_{rec}$ (best close to $RepRatio_{gt}$). RepBias and logDP (best close to zero).

Meth.	Diversity optimization						Item fairness optimization					
	Type	Recall↑	DS↑	RepR	RepBias	mDR↑	Type	Recall↑	logDP	RepR	RepBias	mFR↓
TIFUKNN	N	**0.2130**	0.8102	0.8094	0.5994	0.1054	N	**0.2130**	2.9426	0.8094	0.5994	1.7710
	D	0.1903	0.9869	0.7851	0.5752	0.2059	F	0.1906	1.9240	0.8299	0.6199	1.2720
	R	0.2128	0.8048	0.7792	0.5692	0.1178	R	0.2001	3.2601	**0.5851**	**0.3751**	1.8176
	RD	0.1895	**0.9897**	**0.7522**	**0.5422**	**0.2238**	RF	0.1841	**1.9128**	0.7417	0.5317	**1.2223**
UP-CF@r	N	0.1941	0.7874	0.8644	0.6544	0.0665	N	**0.1941**	3.3743	0.8644	0.6544	2.0144
	D	0.1730	0.9846	0.8346	0.6246	0.1800	F	0.1733	1.9177	0.8608	0.6508	1.2843
	R	**0.1942**	0.7872	0.8638	0.6538	0.0667	R	0.1861	3.6144	**0.5908**	**0.3808**	1.9976
	RD	0.1732	**0.9870**	**0.8318**	**0.6218**	**0.1826**	RF	0.1712	**1.8187**	0.7600	0.5500	**1.1844**
DNNTSP	N	**0.1986**	0.8079	0.6944	0.4844	0.1617	N	**0.1986**	3.5654	0.6944	0.4844	2.0249
	D	0.1891	0.9121	0.6863	0.4763	0.2179	F	0.1972	3.2709	0.7033	0.4933	1.8821
	R	0.1951	0.8050	0.6164	0.4064	0.1993	R	0.1781	3.7944	**0.4217**	**0.2117**	2.0031
	RD	0.1853	**0.9265**	**0.6059**	**0.3959**	**0.2653**	RF	0.1801	**3.2350**	0.4752	0.2652	**1.7501**
Dream	N	0.1210	0.7000	0.0858	-0.1242	0.2879	N	0.1210	8.0904	0.0858	-0.1242	4.1073
	D	0.0989	**1.0000**	0.0764	-0.1336	0.4332	F	0.0888	**2.3133**	0.0617	-0.1483	**1.2308**
	R	**0.1400**	0.6985	**0.2163**	**0.0063**	0.3461	R	**0.1398**	7.3774	**0.2143**	**0.0043**	3.6908
	RD	0.1194	0.9932	0.1984	-0.0116	**0.4908**	RF	0.1219	2.8182	0.2148	0.0048	1.4115
TREx	N	**0.1971**	0.7979	0.8703	0.6603	0.0688	N	0.1971	2.8936	0.8703	0.6603	1.7770
	D	0.1941	0.8243	0.8703	0.6603	0.0820	F	0.1843	**2.1799**	0.8703	0.6603	1.4201
	R	0.1818	0.7251	**0.3980**	**0.1880**	0.2686	R	**0.2018**	3.1900	**0.7496**	**0.5396**	1.8648
	RD	0.1779	**0.8350**	**0.3980**	**0.1880**	**0.3235**	RF	0.1875	2.2036	**0.7496**	**0.5396**	**1.3716**

5

Mask-Prompted Generative Retrieval for Diverse and Exploratory Next Basket Recommendation

Exploratory next basket recommendation seeks to surface novel yet relevant items that support long-term user engagement and item discovery. While existing methods predominantly follow a discriminative formulation, generative recommendation offers an alternative paradigm that produces recommendations directly through sequence generation. Adapting semantic-ID-based generative recommendation to exploratory basket recommendation poses a key challenge: the structural mismatch between autoregressive generation and the order-invariant nature of items within a basket. Therefore, we investigate:

RQ4 *How can we design a semantic-ID-based generative recommendation method for exploratory basket recommendation, and how to improve recommendation diversity under the generative paradigm?*

5.1 Introduction

Exploratory next basket recommendation (NBR) aims to recommend novel items that a user may interact with in a future basket, based on the user's historical basket sequence [61]. Different from repetitive NBR [3], which focuses solely on repeat purchase behaviors, exploratory NBR requires the model to infer potential interests beyond observed interactions and recommend items that are both relevant and novel [63, 65]. Exploratory NBR is widely applicable in retail, grocery, and music platforms [110], and is essential for supporting item discovery and long-term user engagement [18, 105, 114]. Existing studies mostly formulate exploratory NBR as a discriminative prediction problem, where the model scores candidate items and selects the top-ranked ones as the next basket [61]. Although effective, these methods typically rely on item-level scoring

This chapter is based on Y. Liu, Y. Tang, M. Li, C. Xu, M. Aliannejadi, and M. de Rijke. Mask-prompted generative retrieval for diverse and exploratory next basket recommendation. Under review, 2026.

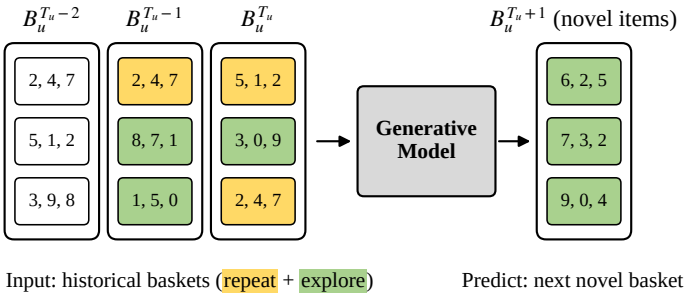


Figure 5.1: Generative exploratory basket recommendation.

objectives and post-processing methods to optimize diversity [77], rather than directly shaping the representation space from which recommendations are produced.

Generative recommendation for NBR. Recently, generative recommendation has emerged as a new paradigm that generates item identifiers directly via autoregressive decoding [99, 120]. This formulation usually follows a two-stage pipeline: it first constructs discrete semantic IDs for items, often through quantization, and then trains a generative recommender to generate target ID tokens conditioned on user history [37, 74]. Owing to its natural ability to capture item correlations within each basket, the generative paradigm has been explored in NBR by autoregressively generating atomic item IDs to form the next basket [108]. However, atomic IDs treat items as opaque tokens and discard their underlying semantic structure. Semantic-ID based generation addresses this by decomposing each item into multiple shared ID tokens, offering a finer-grained item representation and a principled control point over autoregressive generation. We explore the semantic-ID based generative paradigm for exploratory NBR in this work, as illustrated in Figure 5.1.

Challenges. A natural question arises: can the generative recommendation paradigm be adapted to generate baskets of novel items that are both diverse and accurate? Such an adaptation is non-trivial due to two challenges. (i) Current generative recommenders address diversity at the inference stage, e.g., via diverse beam search (DBS) [111], which incurs an accuracy-diversity trade-off [8]. A more fundamental option to intervene lies earlier in the generative recommendation pipeline, at the ID construction stage: quantization-based tokenizers can produce imbalanced ID assignments [120], causing the recommender to learn a skewed target distribution and generate baskets concentrated on a limited region of the ID space. This raises a natural question: *can balancing the distribution over semantic IDs improve recommendation diversity without sacrificing accuracy?* (ii) The autoregressive generation objective does not naturally fit the basket structure. A basket is an unordered yet correlated item set, while semantic-ID based generation predicts ID tokens in an ordered manner. Directly generating the whole basket as a sequence introduces artificial order bias, whereas predicting items independently weakens the modeling of item correlations within the basket. These challenges suggest that exploratory generative NBR requires both diversity-aware semantic ID construction and basket-aware generative modeling.

MPGBR. To address these challenges, we propose mask-prompted generative basket

recommendation (MPGBR) for diverse and exploratory next basket recommendation. The key idea is to improve diversity at the semantic ID construction stage while modelling unordered basket context at the recommender learning stage. First, we construct basket-aware semantic IDs by injecting item co-occurrence signals into item representations, and propose Balanced RQ-KMeans to increase code assignment entropy through optimal transport. This reduces semantic ID concentration and encourages the generative model to cover a broader item space. Second, we design a mask-prompted generative framework that encodes historical baskets with a bidirectional transformer and reconstructs the semantic IDs of masked items with a lightweight decoder. This allows the model to capture item correlations without imposing a fixed generation order over the entire basket.

Experiments. We evaluate MPGBR on three public grocery datasets (TaFeng, Instacart, and Dunnhumby) against strong discriminative and generative baselines. The results show that MPGBR matches the accuracy of strong baselines while advancing the accuracy-diversity Pareto frontier across all datasets. Ablation studies further reveal that balanced code assignment is the primary driver of the diversity gain, confirming representation-level balancing as a key lever for exploratory generative recommendation.

Our contributions. The main contributions of this work are: (i) We introduce exploratory generative next basket recommendation and identify two key challenges: balancing semantic ID distributions to generate diverse baskets, and modeling unordered baskets in a sequential generative framework. (ii) We propose MPGBR, a mask-prompted generative framework that recovers the semantic IDs of masked items from basket context. (iii) We propose Balanced RQ-KMeans, which balances code assignments via optimal transport and thereby improves diversity at the representation stage without sacrificing accuracy. (iv) Extensive experiments on three public grocery datasets show that MPGBR achieves competitive accuracy while delivering substantially better diversity, leading to a more favorable accuracy-diversity trade-off.

5.2 Task Definition

Let a set of users $U = \{u_1, u_2, \ldots, u_v\}$ and a set of items $I = \{i_1, i_2, \ldots, i_m\}$ be given. For each user $u \in U$, there is a sequence of historical basket interactions $H_u = [B_u^1, \ldots, B_u^t, \ldots, B_u^{T_u}]$, where B_u^t represents a set of items that user u purchased at time t, T_u is the length of historical baskets. For a user u, repeat item i_u^{rep} represents an item that user u has purchased in the history, which is defined as $i_u^{rep} \in I_u^{rep} = \bigcup_{t=1}^{T_u} B_u^t$. A novel item is defined as $i_u^{novel} \in I_u^{novel} = I \setminus I_u^{rep}$, i.e., an item that user u has never interacted with. The target of next novel basket recommendation is to predict a basket $B_u^{T_u+1}$, which contains a set of novel items i_u^{novel} tailored to the user's interests.

In generative recommendation, each item i is represented by semantic IDs, i.e., multiple ordered tokens $[s_i^1, \ldots, s_i^L]$, where L is the length of the semantic IDs. We write $\mathcal{S}^l = \{s_i^l\}_{i=1}^m$ for the collection of l-th layer of semantic IDs across all items. Therefore, a basket can be transformed into a set of token sequences $B_u^t = \{[s_i^1, \ldots, s_i^L], \ldots, [s_j^1, \ldots, s_j^L]\}$. Next novel basket recommendation can then be for-

mulated as a sequence-to-sequence generation problem as follows:

$$P(Y \mid H_u) = \prod_{l=1}^{L} P\left(s^l_{T_u+1} \mid H_u, s^1_{T_u+1}, \ldots, s^{l-1}_{T_u+1}\right). \tag{5.1}$$

The top-K items Y with the highest probabilities form the next novel basket $B_u^{T_u+1} = \{Y_1, \ldots, Y_K\}$ recommended to the user.

5.3 MPGBR

We propose MPGBR, a mask-prompted generative framework for diverse and exploratory next basket recommendation. MPGBR consists of two stages: diversity-aware semantic ID construction and mask-prompted generative basket recommendation.

First, we construct diversity-aware item semantic IDs. We use a collaborative semantic modulator to enrich text-based item representations with basket-level collaborative signals, and then apply balanced RQ-KMeans to generate hierarchical semantic IDs with more balanced codeword assignments. These semantic IDs serve as the generation targets of the recommender.

Second, we train a mask-prompted generative recommender to map historical baskets to future-basket items. The encoder captures item correlations from historical baskets, while the decoder autoregressively generates target semantic IDs from encoder-side mask prompts. The generated semantic IDs are then mapped back to items for exploratory next basket recommendation.

5.3.1 Diversity-aware item tokenization

Collaborative Semantic Modulator

Pre-trained language models encode rich textual semantics for items, but they do not explicitly capture basket-level co-occurrence patterns that are important for next basket recommendation. To enrich item tokenization with collaborative signals while preserving textual semantics, we propose a collaborative semantic modulator (CSM), which treats the text embedding as a semantic anchor and injects co-occurrence information through a bounded residual modulation. CSM consists of two components: bounded collaborative modulation, which constructs the final item representation, and co-occurrence-guided semantic alignment, which supervises the modulation using basket co-occurrence signals.

Bounded collaborative modulation. For each item i, we first use a pre-trained language model to encode its content information into an initial representation $\mathbf{z}_i$. As shown in Figure 5.2, two parallel transformations are applied to $\mathbf{z}_i$. A linear projection produces the text-based semantic representation: $\mathbf{z}_i^{\text{text}} = \mathbf{W}_{\text{text}}\mathbf{z}_i + \mathbf{b}_{\text{text}}$, while a multilayer perceptron network (MLP) produces a collaborative residual shift: $\mathbf{z}_i^{\text{cf}} = \text{MLP}_{\text{cf}}(\mathbf{z}_i)$.

The final modulated item representation is defined as: $\hat{\mathbf{z}}_i = \mathbf{z}_i^{\text{text}} + \gamma \cdot \mathbf{z}_i^{\text{cf}}$, where $\gamma \ll 1$ controls the magnitude of the collaborative shift. This bounded residual design allows collaborative signals to adjust the item representation while keeping the representation anchored in textual semantics.

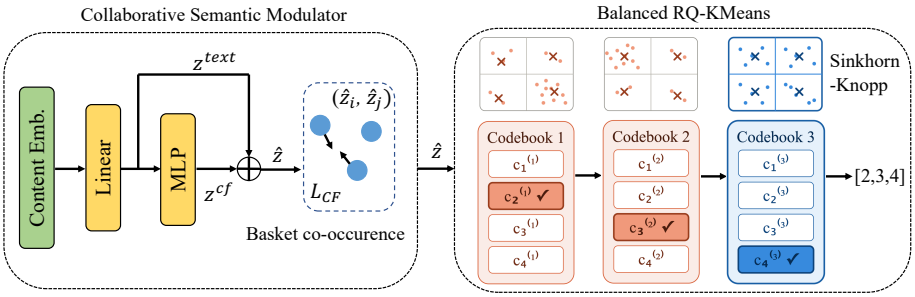


Figure 5.2: Diversity-aware Item Tokenization. Item content representations are first modulated with basket-level collaborative signals through a Collaborative Semantic Modulator, then quantized into hierarchical semantic IDs by Balanced RQ-KMeans.

Co-occurrence-guided semantic alignment. To guide the collaborative modulation, we use basket co-occurrence as supervision. The intuition is that items frequently appearing in the same basket should have closer representations in the modulated semantic space. Given a batch of baskets $\{B_i\}_{i=1}^{N}$ sampled from historical interactions, we construct a co-occurrence matrix and treat item pairs appearing in the same basket as positive pairs. Negative pairs are obtained through in-batch sampling, where item pairs that do not co-occur in any basket within the batch are treated as negatives.

Following [59], we optimize the modulated representations with the Sigmoid Contrastive Loss:

$$\mathcal{L}_{\text{CF}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{Y_i}\sum_{j=1}^{N}\log\left(1+e^{y_{ij}(-t\langle\hat{\mathbf{z}}_i,\hat{\mathbf{z}}_j\rangle+b)}\right), \tag{5.2}$$

where $\langle\hat{\mathbf{z}}_i,\hat{\mathbf{z}}_j\rangle$ denotes the cosine similarity between two modulated item representations. The label $y_{ij}=1$ indicates a positive co-occurring pair, while $y_{ij}=-1$ indicates a negative pair. Here, N is the batch size and Y_i denotes the number of positive pairs associated with item i. The temperature is parameterized as $t=e^{t'}$, where both t' and the bias term b are learnable parameters.

The resulting representation $\hat{\mathbf{z}}_i$ is then used for semantic ID construction, allowing the generated IDs to reflect both textual semantics and basket-level collaborative signals.

Balanced RQ-KMeans

Given the fused item embeddings, our goal is to construct hierarchical semantic IDs that preserve item semantics while encouraging diverse codeword usage. However, existing RQ-based tokenizers, such as RQ-KMeans [27], mainly optimize reconstruction quality and can produce skewed codeword assignments, which limits the diversity of generated IDs. To address this issue, we propose balanced RQ-KMeans, which introduces a balance-aware assignment step into residual quantization. As shown in the right-hand side of Figure 5.2, instead of assigning each item solely to the nearest codeword, we jointly consider residual distance and codeword usage balance. We formulate

this assignment as an entropy-regularized optimal transport problem, where residual distances define the assignment cost and balanced marginals encourage more uniform codeword usage. We solve this problem using the Sinkhorn-Knopp algorithm [24], which efficiently computes balanced soft assignments.

Specifically, at each quantization layer, balanced RQ-KMeans consists of three steps: codebook construction, code assignment, and residual update. The code assignment step performs balance-aware assignment at selected layers and standard nearest-centroid assignment at the others.

Codebook construction. Following [27], we initialize the residual set as the fused item embeddings: $\mathcal{R}^{(1)} = \{\hat{\mathbf{z}}_i \in \mathbb{R}^d \mid \forall \text{ item } i\}$, where d is the embedding dimension. At layer l, we construct a codebook by applying Mini-Batch K-means to the current residuals: $\mathcal{C}^{(l)} = \text{K-means}(\mathcal{R}^{(l)}, N)$, where $\mathcal{C}^{(l)} = \{\mathbf{c}_k^{(l)} \in \mathbb{R}^d \mid k = 1, \ldots, N\}$ and N is the codebook size.

Balance-aware assignment. For selected layers $l \in \mathcal{L}_{\text{bal}} \subseteq \{1, \ldots, L\}$, we compute item-to-codeword assignments through an optimal transport problem. Given the residual set $\mathcal{R}^{(l)}$ and the codebook $\mathcal{C}^{(l)}$, we solve for a soft assignment matrix $\mathbf{Q}^{(l)} \in \mathbb{R}_+^{|B| \times N}$, where $\mathbf{Q}_{i,k}^{(l)} = q(s_i^l = k \mid \mathcal{R}_i^{(l)})$ denotes the probability of assigning item i to the k-th centroid:

$$\begin{aligned} \min_{\mathbf{Q}^{(l)}} \quad & \sum_{i=1}^{|B|} \sum_{k=1}^{N} \mathbf{Q}_{i,k}^{(l)} \left\| \mathcal{R}_i^{(l)} - \mathbf{c}_k^{(l)} \right\|_2^2 \\ \text{such that } & \sum_{k=1}^{N} \mathbf{Q}_{i,k}^{(l)} = 1, \quad \forall i, \\ & \sum_{i=1}^{|B|} \mathbf{Q}_{i,k}^{(l)} = \frac{|B|}{N}, \quad \forall k. \end{aligned} \tag{5.3}$$

The first constraint ensures that each item is fully assigned, while the second enforces a uniform marginal over codewords, so that each centroid receives an equal share of items. We solve Eqn. 5.3 efficiently using the Sinkhorn-Knopp algorithm [24], where $\epsilon > 0$ controls the sharpness of the transport plan. After solving Eqn. 5.3, the discrete index for item i at layer l is obtained by: $s_i^l = \arg\max_k \mathbf{Q}_{i,k}^{(l)}$.

For layers $l \notin \mathcal{L}_{\text{bal}}$, we use the nearest-centroid assignment as in RQ-KMeans [27]:

$$s_i^l = \arg\min_k \left\| \mathcal{R}_i^{(l)} - \mathbf{c}_k^{(l)} \right\|_2. \tag{5.4}$$

Residual update. After obtaining the codeword index s_i^l, the residual of item i is updated as: $\mathcal{R}_i^{(l+1)} = \mathcal{R}_i^{(l)} - \mathbf{c}_{s_i^l}^{(l)}$. Repeating this process over L layers yields the final hierarchical semantic ID: $\mathbf{s}_i = [s_i^1, \ldots, s_i^L]$. At balanced layers, the codebook $\mathcal{C}^{(l)}$ and the assignment $\mathbf{Q}^{(l)}$ are coupled; we therefore alternate between K-means centroid updates and Sinkhorn-based assignments until convergence. Let $\mathcal{S}^l = \{s_i^l\}_{i=1}^m$ denote the collection of l-th layer codes across all items. At balanced layers, the distribution of $\mathcal{S}^l$ should become more uniform. Unlike prior work [59, 144, 150] that applies Sinkhorn-Knopp for collision resolution, our approach targets downstream

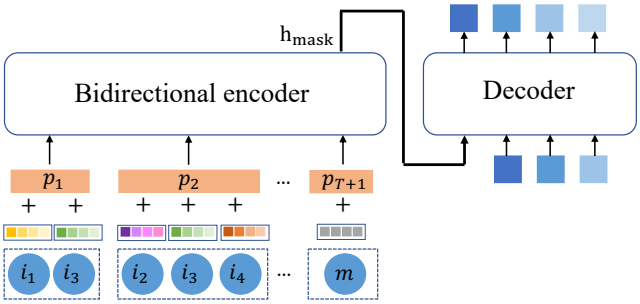


Figure 5.3: MPGBR architecture.

recommendation diversity, achieved through well-balanced semantic IDs. Finally, we append an additional code to each semantic ID to guarantee uniqueness across items.

5.3.2 Mask-prompted generative basket recommendation

Once the semantic IDs of items have been obtained, we train a generative recommender to model the mapping from a user's historical basket sequence to the items in the future basket. Instead of directly predicting item indices, the model generates the semantic ID of each target item conditioned on the user's historical interactions.

Model architecture. As illustrated in Figure 5.3, MPGBR adopts a transformer encoder-decoder architecture. The bidirectional encoder models both intra-basket item correlations and cross-basket interaction patterns from the user's historical baskets. The lightweight autoregressive decoder then uses the encoder-side mask representation as a prompt to generate the semantic ID of the target item.

Input encoding. For each item i, we first represent its semantic ID $[s_i^1, \ldots, s_i^L]$ by summing the embeddings of all ID tokens: $\mathbf{E}_i = \sum_{j=1}^{L} \mathbf{e}_{s_i^j}$, where $\mathbf{e}_{s_i^j}$ denotes the embedding of the j-th token in the semantic ID of item i. Following [61], we use a shared basket-level position embedding $\mathbf{p}_t$ for all items in the t-th basket. The encoder input for item i in basket t is defined as: $\mathbf{x}_{i,t} = \mathbf{E}_i + \mathbf{p}_t$. This design preserves the sequential order among baskets while treating items within the same basket as an unordered set.

Mask-prompted ID generation. During training, selected items in the basket sequence are replaced with a special $[\text{mask}]$ token. Given the masked sequence, the encoder produces contextualized hidden states; the hidden state at each mask position is then linearly transformed into a context-aware prompt for semantic ID generation:

$$\mathbf{h}_m = Lin(\text{Encoder}(\mathbf{x}_1, \ldots, \mathbf{x}_{m-1}, [\text{mask}], \mathbf{x}_{m+1}, \ldots, \mathbf{x}_n)). \tag{5.5}$$

Following [59], we use a lightweight autoregressive decoder to generate the semantic ID of the masked item:

$$P(\mathbf{s}_i \mid \mathbf{h}_m) = \prod_{j=1}^{L} P(s_i^j \mid \mathbf{h}_m, s_i^{<j}), \tag{5.6}$$

where $\mathbf{s}_i = [s_i^1, \ldots, s_i^L]$ denotes the semantic ID of item i, and $s_i^{<j}$ denotes the tokens before the j-th token. The recommendation loss is defined as:

$$\mathcal{L}_{\mathrm{REC}} = -\sum_{j=1}^{L} \log P(s_i^j \mid \mathbf{h}_m, s_i^{<j}). \tag{5.7}$$

Training and inference. Following [61], we adopt a pretraining-finetuning training paradigm. During pretraining, a subset of items in the historical basket sequence is masked, encouraging the model to learn item-item correlations from observed interactions. Given a user's basket sequence, we randomly sample a fraction ρ of items from the user's interacted item set and mask all their occurrences in the basket sequence.

During finetuning, the entire target basket is masked, which aligns the training objective with the next basket recommendation task. To enrich within-basket interactions, we further apply the swapping strategy during pretraining [61]. We randomly select items with a swap ratio α and move them to other baskets, where the swap hop β controls the basket distance of the swap.

At inference time, we place $[\mathrm{mask}]$ tokens at the target basket positions and use their hidden states as prompts for semantic ID generation. The decoder generates candidate semantic IDs autoregressively, and the generated IDs are mapped back to items through the semantic ID table. We rank candidate novel items from I_u^{novel} by the log-likelihood of their generated semantic IDs and select the top-K as the recommended basket.

5.4 Experimental Setup

Datasets. Following previous NBR studies [61, 63, 65, 75], we evaluate MPGBR on three publicly available grocery shopping datasets: TaFeng [109], Dunnhumby [29], and Instacart [48]. The item content information for the three datasets is as follows:

- **TaFeng**: no item content information is available.
- **Dunnhumby**: hierarchical categories: department, commodity_desc, sub_commodity_desc.
- **Instacart**: department, aisle, and product name.

Following Ariannezhad et al. [3], we remove users with fewer than three baskets and items purchased fewer than five times from raw datasets. For Instacart, whose size caused memory issues for several baselines, we sample 20,000 users before filtering, as in [87]. We adopt the data-splitting protocol of [3, 33, 87]. Following Li et al. [63], all baskets except the last are used for training; we cap the training history at the 50 most recent baskets per user. The held-out last baskets are randomly split into validation and test sets in equal proportions. Table 5.1 summarizes the statistics of the three datasets after preprocessing. The average $RepRatio_{gt}$, defined as in [63], measures the average proportion of repeat items in the ground-truth baskets.

NBR baselines. We select and examine six discriminative NBR methods and two generative methods: **GTop** recommends the top k most popular items in the dataset as the next

Table 5.1: Statistics of datasets after preprocessing.

Dataset	#User	#Item	#Basket	Avg. #B/U	Avg. #I/B	Avg. $RepRatio_{gt}$	#Cate -gory
TaFeng	10,182	15,023	82,387	8.09	6.14	0.21	1669
Dunnhumby	2,482	37,161	107,152	43.17	10.07	0.43	34
Instacart	19,210	29,398	305,582	15.91	10.06	0.60	21

basket for users. **Dream** derives basket representations through pooling and leverages recurrent neural network (RNN) to capture sequential user behavior [137]. **DNNTSP** employs graph neural networks (GNNs) and self-attention to encode item-to-item relations across baskets while modeling temporal dependencies [138]. **UP-CF@r** integrates recency-aware user-wise popularity with collaborative filtering, and account for recent shopping behavior [33]. **TIFUKNN** captures temporal dynamics in users' basket histories via a KNN-based approach modeling the personal item frequency (PIF) [45]. **BTBR** adopts a bidirectional transformer encoder to learn item-item correlations within and across baskets for exploratory basket recommendation [61].

We exclude [108, 147] as their code is not publicly available. In their place, we adapt generative sequential recommendation methods as additional baselines, since they likewise operate on item sequences. Specifically, we flatten the historical baskets into an item sequence and split the target basket into multiple individual items. **TIGER** proposes RQ-VAE to convert item content embeddings to semantic IDs, and a transformer encoder-decoder architecture is employed for next item prediction [99]. **LETTER** enhances RQ-VAE by injecting collaborative signals and alleviating code assignment bias [120]. As TaFeng lacks item text information, we introduce $\textbf{MPGBR}_{CF}$, a variant that constructs semantic IDs from collaborative filtering (CF) embeddings extracted from a pre-trained BERT4Rec [106].

Diversity baseline. We adopt diverse beam search (DBS), which divides the beam into groups and applies a diversity penalty across them to encourage distinct outputs [111].

Metrics. We evaluate along two complementary axes: accuracy and diversity. Accuracy is measured by Recall and normalized discounted cumulative gain (NDCG), where Recall captures the proportion of ground-truth items recovered in the top-K recommendations and NDCG additionally rewards ranking ground-truth items higher in the list. Diversity is measured by intra-list diversity (ILD), the average pairwise category dissimilarity between recommended items (ranging in $[0, \sqrt{2}]$), and diversity score (DS), the number of distinct categories in the recommended basket normalized by basket size. All metrics are reported at cutoffs $K \in \{10, 20\}$.

Implementation details. *Item embedding construction.* For Instacart and Dunnhumby, we use the pretrained Sentence-T5-XL[1] to generate item content embeddings. For TaFeng, where item textual metadata is limited, we extract $d = 128$ item embeddings from BERT4Rec. All hyperparameters are selected by grid search on the validation set.

Semantic ID construction. For CSM, the weight γ is selected from $\{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. In the contrastive loss, we initialize the temperature and bias as $t' = 2$

[1] https://huggingface.co/sentence-transformers/sentence-t5-xl

and $b = -8$, respectively. For balanced RQ-KMeans, each codebook contains 256 codewords. We use three codebooks and append one additional token, resulting in a final semantic ID length of $L = 4$. The entropy regularization coefficient ϵ in the Sinkhorn-Knopp algorithm is tuned from $\{0.01, 0.03, 0.05, 0.07, 0.09, 0.1\}$.

Model configuration and training. For MPGBR, both the transformer encoder and decoder contain 2 layers with 8 attention heads, and the token embedding size is 128. We optimize the model using Adam with a learning rate of 0.001 and a batch size of 128. The model is trained for 50 epochs on Instacart and TaFeng, and 200 epochs on Dunnhumby. The mask ratio ρ is selected from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, the swap ratio α from $\{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$, and the swap hop β from $\{1, 3, 5, 7, 9\}$. We implement the model in PyTorch and train it on an NVIDIA H100 GPU. For all methods, we run five random seeds and report the average results.

5.5 Results

To evaluate the accuracy and diversity performance of MPGBR, we focus on five sub-research questions:

(RQ4.1) What is the accuracy and diversity performance of MPGBR compared with NBR baselines?

(RQ4.2) How does each component of our proposed MPGBR framework contribute to its overall effectiveness?

(RQ4.3) At which semantic ID level is the balancing algorithm most effective for improving recommendation diversity?

(RQ4.4) How does the diversity-accuracy trade-off behave when comparing balanced RQ-KMeans with diverse beam search?

(RQ4.5) How are semantic IDs distributed across codebook levels under balanced RQ-KMeans quantization, and to what extent do distributional properties such as codewords assignment entropy account for diversity gains in downstream generative recommendation?

Accuracy performance. To answer **(RQ4.1)**, we experiment with MPGBR and two categories of baselines, namely, six discriminative recommendation methods and two generative recommendation methods, across three datasets (TaFeng, Dunnhumby, Instacart). We evaluate all methods using Recall@K and NDCG@K; see Table 5.2. We also report on MPGBR$_{CF}$, a variant whose semantic IDs are generated solely from CF embeddings extracted from BERT4Rec.

We make the following observations: (i) MPGBR$_{CF}$ consistently achieves either the best or second-best performance across all metrics and datasets. MPGBR$_{CF}$ consistently outperforms the strongest baseline BTBR on the TaFeng dataset and remains competitive on Dunnhumby and Instacart. This demonstrates the overall effectiveness of our proposed approach for exploratory basket recommendation. (ii) The generative recommendation models (TIGER and LETTER) perform poorly across all three

datasets. This highlights a critical limitation: generative architectures designed for sequential recommendation do not readily transfer to the next-basket setting, which requires predicting a set of co-occurring items rather than a single next item. This finding motivates the design of MPGBR, which is tailored to basket-level prediction. (iii) While MPGBR$_{CF}$ achieves marginally better results than MPGBR on Dunnhumby and Instacart, this improvement should be interpreted with caution. We attribute the gap primarily to the limited content information in Dunnhumby, where items are described only by coarse-grained hierarchical categories without product names, weakening the content-based representation. (iv) Semantic IDs should be grounded in item content rather than CF embeddings alone, for two reasons: (i) **coverage:** MPGBR$_{CF}$ derives its semantic IDs from CF embeddings learned by BERT4Rec, which are only available for items observed during training. In practice, an item tokenizer should generate semantic IDs for the entire item catalog, including items that appear only in the test set; and (ii) **generalization:** content-grounded semantic IDs generalize to cold-start items, whereas CF-grounded IDs do not.

Accuracy-diversity trade-off analysis. To assess whether MPGBR[2] achieves a favorable balance between recommendation accuracy and diversity, we plot Recall@K against ILD@K for $K \in \{10, 20\}$ on all three datasets. The results are shown in Figure 5.4. A method positioned toward the upper-right corner of each subplot is preferable, as it simultaneously achieves high accuracy and high diversity.

Based on Figure 5.4, we observe the following: (i) Across the six evaluation settings (3 datasets × 2 cutoffs), MPGBR lies on the Pareto frontier in 5 out of 6 cases, demonstrating a consistently strong accuracy-diversity trade-off. On TaFeng, MPGBR achieves the highest Recall and a competitive ILD. On Dunnhumby, MPGBR and the baselines exhibit a clear accuracy-diversity trade-off, with no single method dominating on both axes. On Instacart, MPGBR achieves competitive Recall (slightly below BTBR) and higher diversity than BTBR, placing it on the Pareto frontier. (ii) ILD ranges differ substantially across datasets due to catalog structure rather than methodology. TaFeng's 1,669 item categories (vs. 21 in Instacart and 34 in Dunnhumby) inflate diversity scores uniformly and compress the gap between methods. GTop, the simplest baseline, confirms this: its ILD@20 exceeds 1.35 on TaFeng (limited headroom for improvement) but is near 0.1 on Instacart (substantial headroom). These structural differences account for the modest diversity gains observed on TaFeng compared to Instacart.

These results show that MPGBR lies on the Pareto frontier of the accuracy-diversity trade-off in 5 out of 6 settings, delivering competitive or superior accuracy without sacrificing diversity, validating the effectiveness of the balanced RQ-KMeans item tokenizer and MPGBR framework.

Ablation study. To answer **(RQ4.2)**, we conduct an ablation study by systematically removing or disabling individual components of MPGBR and evaluating the resulting variants on all three datasets. We consider the following variants:

- **w/o fine-tune**: removes the fine-tuning stage, focusing solely on learning item-item correlations during pre-training.

[2]Hereafter, we use MPGBR on Instacart and the variant MPGBR$_{CF}$ on Dunnhumby and Tafeng, but refer to both as MPGBR for brevity.

Table 5.2: Accuracy comparison of MPGBR with baseline methods across three datasets. The best result per metric is highlighted in bold, while the second best is underlined. $*$ indicates that the improvement of the best method over the second best is statistically significant under a paired t-test ($p < 0.05$).

Method	**TaFeng**				**Dunnhumby**				**Instacart**			
	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
Discriminative methods												
GTop	0.0583	0.0803	0.0604	0.0678	0.0100	0.0078	0.0146	0.0099	0.0322	0.0512	0.0251	0.0328
Dream	0.0727	0.0978	0.0696	0.0791	0.0186	0.0283	0.0150	0.0194	0.0401	0.0626	0.0329	0.0420
DNNTSP	0.0787	0.1036	0.0726	0.0816	0.0174	0.0267	0.0147	0.0188	0.0414	0.0633	0.0338	0.0429
UPCF	0.0739	0.0953	0.0687	0.0771	0.0196	0.0294	**0.0162**	0.0208	0.0444	0.0695	0.0366	0.0466
TIFUKNN	0.0807	0.1072	0.0760	0.0856	0.0191	0.0289	0.0140	0.0177	0.0446	0.0689	0.0362	0.0460
BTBR	0.0831	0.1078	0.0747	0.0839	0.0189	**0.0330***	0.0159	**0.0215**	**0.0591***	**0.0906***	**0.0472***	**0.0596***
Generative methods												
TIGER	0.0716	0.0916	0.0688	0.0765	0.0153	0.0251	0.0127	0.0171	0.0470	0.0735	0.0376	0.0478
LETTER	0.0659	0.0900	0.0631	0.0719	0.0151	0.0251	0.0125	0.0169	0.0529	0.0837	0.0422	0.0542
MPGBR$_{CF}$	**0.0860***	**0.1101**	**0.0785***	**0.0878***	**0.0206**	0.0317	**0.0162**	0.0212	0.0545	0.0847	0.0438	0.0556
MPGBR	[——— *Not applicable* ———]				0.0184	0.0299	0.0143	0.0191	0.0534	0.0842	0.0421	0.0541

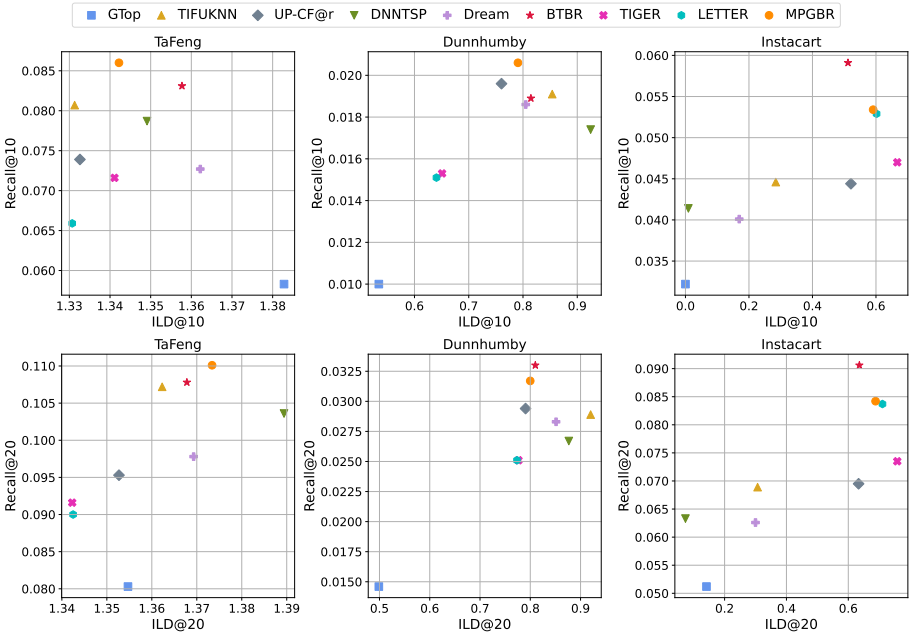


Figure 5.4: Accuracy and diversity evaluation of MPGBR and baselines. Method in the upper-right corner is better (higher accuracy and diversity). The y-axis is scaled up for better visibility.

- **w/o pre-train**: removes the pre-training stage, training the model directly on the task of predicting items in the next basket.
- **w/o CSM**: removes the collaborative modulation in CSM, directly using item content embeddings to generate semantic IDs.
- **w/o Balanced**: replaces the balance-aware assignment with standard RQ-KMeans.

The pre-train/fine-tune stages are designed to improve accuracy, while CSM and the balanced RQ-KMeans target recommendation diversity. We report Recall@10 (R@10), NDCG@10 (N@10), ILD@10, and DS@10 to jointly assess how these components affect accuracy and diversity. Results are presented in Table 5.3. Missing entries indicate 'w/o CSM' is not applicable to Dunnhumby and TaFeng, where semantic IDs are generated solely from CF embeddings.

From Table 5.3, we observe that the full model achieves strong overall performance, while the pattern of metric improvements and degradations confirms that each component plays its designated role: (i) Pre-training and fine-tuning jointly drive accuracy, with their relative importance varying across datasets. Removing pre-training (Row 3) causes the largest accuracy drop on Instacart (R@10: $0.0534 \rightarrow 0.0464$, -13.1%) and Dunnhumby (R@10: $0.0206 \rightarrow 0.0162$, -21.4%), whereas on TaFeng removing fine-tuning leads to the larger drop (R@10: $0.0860 \rightarrow 0.0730$, -15.1%). This pattern is consistent with the dataset characteristics: TaFeng has 79% exploratory items in

Table 5.3: Ablation study on accuracy and diversity performance across three datasets.

No.	Variant	Instacart				Dunnhumby				TaFeng			
		R@10	N@10	ILD@10	DS@10	R@10	N@10	ILD@10	DS@10	R@10	N@10	ILD@10	DS@10
1	MPGBR	0.0534	0.0421	0.5911	0.2876	**0.0206**	**0.0162**	0.7907	0.3572	**0.0860**	**0.0785**	**1.3422**	**0.7957**
2	w/o fine-tune	0.0514	0.0412	0.4833	0.2709	0.0182	0.0149	**0.8168**	**0.3757**	0.0730	0.0681	1.3316	0.7658
3	w/o pre-train	0.0464	0.0365	**0.6445**	**0.2884**	0.0162	0.0131	0.6182	0.3054	0.0826	0.0751	1.3290	0.7773
4	w/o CSM	0.0528	0.0420	0.5574	0.2773	[			*Not applicable*				]
5	w/o Balanced	**0.0536**	**0.0422**	0.5539	0.2760	0.0192	0.0146	0.7804	0.3537	0.0857	0.0782	1.3397	0.7778

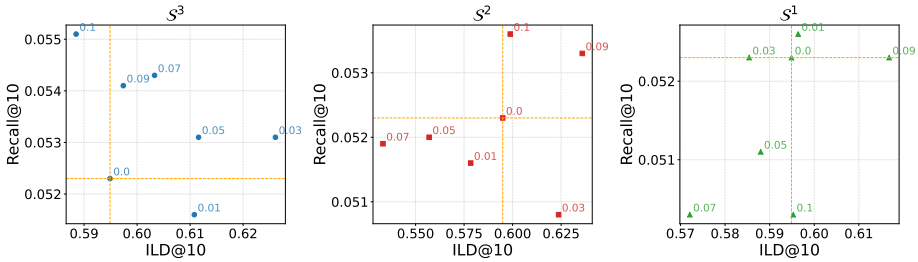


Figure 5.5: Effect of balance-aware assignment at different semantic ID layers ($\mathcal{S}^1$, $\mathcal{S}^2$, $\mathcal{S}^3$) on Instacart. Applying balancing at the deepest layer, $\mathcal{S}^3$, achieves the best accuracy-diversity trade-off. The y-axis is scaled for visibility.

ground-truth baskets, providing rich supervised signals that benefit the fine-tuning stage for exploratory recommendation. In contrast, Instacart and Dunnhumby are more repetition-dominated, making the foundational item representations learned during pre-training more critical. Moreover, removing accuracy-oriented components has an unpredictable impact on diversity, occasionally even improving it. For example, *w/o pre-train* (Row 3) achieves higher ILD@10 and DS@10 on Instacart, while *w/o fine–tune* (Row 2) attains higher diversity scores on Dunnhumby. These instances illustrate a classic accuracy-diversity trade-off: removing accuracy-focused components can incidentally inflate diversity, but at the cost of substantially lower Recall and NDCG. (ii) The CSM component substantially boosts diversity with negligible accuracy cost. On Instacart, removing CSM (Row 4) reduces ILD@10 by 5.7% ($0.5911 \rightarrow 0.5574$) and DS@10 by 3.6% ($0.2876 \rightarrow 0.2773$). We attribute this to CSM's collaborative modulation of content embeddings: by injecting collaborative signals into the original content representations, CSM produces more discriminative fused embeddings and, in turn, a more uniformly distributed semantic ID space, ultimately enhancing recommendation diversity. (iii) Balanced RQ-KMeans consistently improves diversity with virtually no accuracy cost. Removing it (Row 5) reduces ILD@10 and DS@10 on all three datasets, while Recall and NDCG remain comparable to the full model. On Instacart, removing balanced assignment (Row 5) reduces ILD@10 by 6.3% ($0.5911 \rightarrow 0.5539$). This pattern reveals that Balanced RQ-KMeans delivers meaningful diversity gains at minimal accuracy cost, validating its role as a diversity-oriented component in MPGBR.

Our ablation results confirm that each component of MPGBR plays a complementary role: the pre-train/fine-tune stages jointly drive accuracy, while CSM and the balance-aware assignment are essential for achieving high diversity. Together, they enable MPGBR to strike the most favorable balance between accuracy and diversity.

ID levels analysis. To answer **(RQ4.3)**, we conduct experiments on the Instacart dataset by applying balance-aware assignment at each of the three semantic ID layers ($\mathcal{S}^1$, $\mathcal{S}^2$, $\mathcal{S}^3$) independently. For each layer, we sweep ϵ over $\{0, 0.01, 0.03, 0.05, 0.07, 0.09, 0.1\}$ and measure Recall@10 and ILD@10. The results are presented in Figure 5.5, where the orange dashed lines mark the performance of the baseline ($\epsilon = 0$) for reference.

We can observe that applying balance-aware assignment at the third layer of semantic IDs ($\mathcal{S}^3$) yields the most favorable accuracy-diversity trade-off. On $\mathcal{S}^3$, the assignment achieves greater diversity gain over the baseline while incurring the smallest accuracy

drop; several configurations even improve Recall@10 above the baseline. In contrast, applying it at $\mathcal{S}^1$ or $\mathcal{S}^2$ improves diversity in a few configurations but at the cost of larger Recall@10 drops.

We attribute this layer-dependent pattern to the hierarchical structure of semantic IDs: the earlier layers ($\mathcal{S}^1$, $\mathcal{S}^2$) encode main semantic information, so enforcing balance at these layers disrupts the representations reconstruction and harms accuracy. By the last layer ($\mathcal{S}^3$), the semantic space is already sparse and fine-grained, leaving sufficient room for balance-aware assignment to redistribute items more uniformly, thereby improving diversity without compromising Recall. The results at $\mathcal{S}^3$ are also the most stable across hyperparameter values, reducing sensitivity to coefficient tuning in deployment. We therefore apply balance-aware assignment only at the semantic ID layer $\mathcal{S}^3$ in the final MPGBR configuration.

Comparison of balanced RQ-KMeans and diverse beam search (DBS). To answer **(RQ4.4)**, we compare the two methods on all three datasets by sweeping their respective hyperparameters and measuring Recall@20 and ILD@20 at each setting. Specifically, we vary the diversity penalty λ of DBS over $\{0.01, 0.025, 0.05, 0.075, 0.1, 0.125, 0.25, 0.5, 1.0, 1.5, 2.0, 3.0\}$. and the coefficient ϵ in the Sinkhorn-Knopp algorithm is tuned over $\{0.01, 0.03, 0.05, 0.07, 0.09, 0.1\}$. Figure 5.6 presents the results; each node represents one hyperparameter setting, and the red star marks the baseline using neither algorithm.

From these results, we draw the following observations: (i) DBS consistently sacrifices Recall@20 for diversity across all three datasets. On Instacart, DBS pushes ILD@20 from $\sim$0.70 up to $\sim$0.93, but at the cost of Recall@20 dropping from $\sim$0.083 (baseline) to as low as $\sim$0.065. A similar trend appears on TaFeng. However, DBS appears to fail on Dunnhumby, where ILD@20 even decreases slightly relative to the baseline. We attribute this to a fundamental limitation of DBS: by encouraging diversity at the token level during decoding, DBS does not reliably translate into meaningful category-level diversity in the generated baskets, while still incurring an accuracy penalty. (ii) Balanced RQ-KMeans achieves diversity gains while preserving Recall@20. Across all three datasets, there exist configurations where balanced RQ-KMeans matches or exceeds the baseline on Recall@20, even on the challenging Dunnhumby dataset, where DBS fails. Moreover, several configurations on Instacart and TaFeng improve both Recall@20 and ILD@20 over the baseline simultaneously. Notably, accuracy remains robust to the shift in semantic ID distribution induced by balanced RQ-KMeans, suggesting that addressing diversity at the representation level avoids the accuracy-diversity trade-off commonly observed at the inference stage. (iii) The contrast reveals a fundamental difference between the two approaches. DBS operates at the decoding stage and promotes diversity at the token level, which does not strictly correspond to category-level diversity. In contrast, balanced RQ-KMeans enforces a uniform distribution over semantic IDs during tokenization, ensuring that diversity is grounded in the item representation space itself. This explains why representation-level diversification delivers diversity gains without compromising recommendation accuracy, while inference-level diversification (DBS) suffers a strict accuracy cost.

Visualization of semantic ID distribution and explanation. To answer **(RQ4.5)**, we visualize the mapping between item categories and each of the three semantic ID

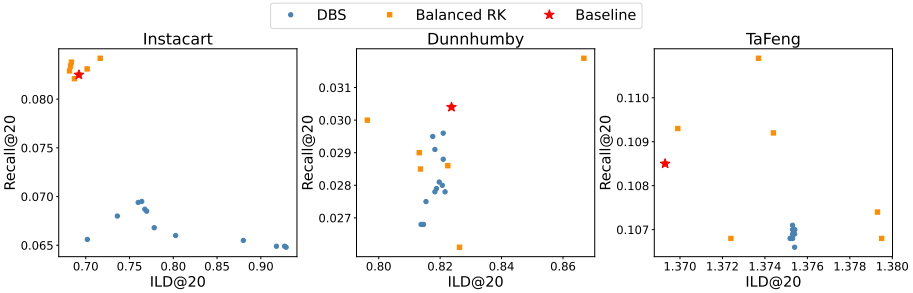


Figure 5.6: Comparison of Balanced RQ-KMeans and Diverse Beam Search across their hyperparameter search spaces. Each point represents one hyperparameter configuration; the red star marks the baseline without either algorithm applied. Points in the upper-right region indicate Pareto-improvement over the baseline.

codebooks ($\mathcal{C}^{(1)}$, $\mathcal{C}^{(2)}$, $\mathcal{C}^{(3)}$) on the Instacart dataset. Specifically, for each layer, we compute the proportion of items in each category that fall on each discrete semantic ID value (0–255), then normalize by the uniform baseline of $1/256$. The resulting relative deviation, $(\text{proportion} - \text{baseline})/\text{baseline}$, is plotted as a heatmap with a diverging red–white–blue colorscale: red indicates concentration above the uniform baseline, white indicates a near-uniform distribution, and blue indicates avoidance. We sample five representative departments from Instacart: meat & seafood (cat12), babies (cat18), bakery (cat3), beverages (cat7), and snacks (cat19). The results are presented in Figure 5.7, which gives rise to the following observations: (i) Categories exhibit the sharpest concentration on the $\mathcal{C}^{(1)}$ codebook. The $\mathcal{C}^{(1)}$ heatmap is dominated by isolated red bands against a dark blue background, indicating that each category sharply prefers a small number of specific semantic ID values. This confirms that the first codebook layer captures coarse-grained semantic structure, where items belonging to the same category cluster onto a few shared codes. (ii) The concentration weakens progressively from $\mathcal{C}^{(1)}$ to $\mathcal{C}^{(3)}$. Moving from $\mathcal{C}^{(1)}$ to $\mathcal{C}^{(2)}$ to $\mathcal{C}^{(3)}$, the heatmaps shift from sparse, high-contrast bands to increasingly diffuse and uniformly light-colored patterns. By $\mathcal{C}^{(3)}$, the relative deviations are close to zero across most semantic ID values, indicating that items within the same category are spread approximately uniformly over the deepest codebook. This validates that the deeper layers encode fine-grained, category-agnostic distinctions. (iii) The observed hierarchy directly justifies our design choice for the balanced RQ-KMeans. The progressive flattening from $\mathcal{C}^{(1)}$ to $\mathcal{C}^{(3)}$ aligns with the layer-ablation finding (Figure 5.5): applying balance-aware assignment at the deepest codebook ($\mathcal{C}^{(3)}$) is most effective because the semantic space is already nearly uniform there, leaving room to redistribute items without disrupting the coarse-grained category structure encoded in $\mathcal{C}^{(1)}$ and $\mathcal{C}^{(2)}$. Conversely, enforcing balance at $\mathcal{C}^{(1)}$ would disrupt the coarse-grained category groupings encoded there, which the model relies on for accurate recommendation.

To understand *why* flattening the distribution of the deepest layer of semantic ID ($\mathcal{S}^3$) translates into improved recommendation diversity, we conduct a fine-grained study on the Instacart dataset. We vary the balance coefficient ϵ on $\mathcal{S}^3$ and measure Recall@20, ILD@20, and the entropy of the codewords assignment. Figure 5.8 shows

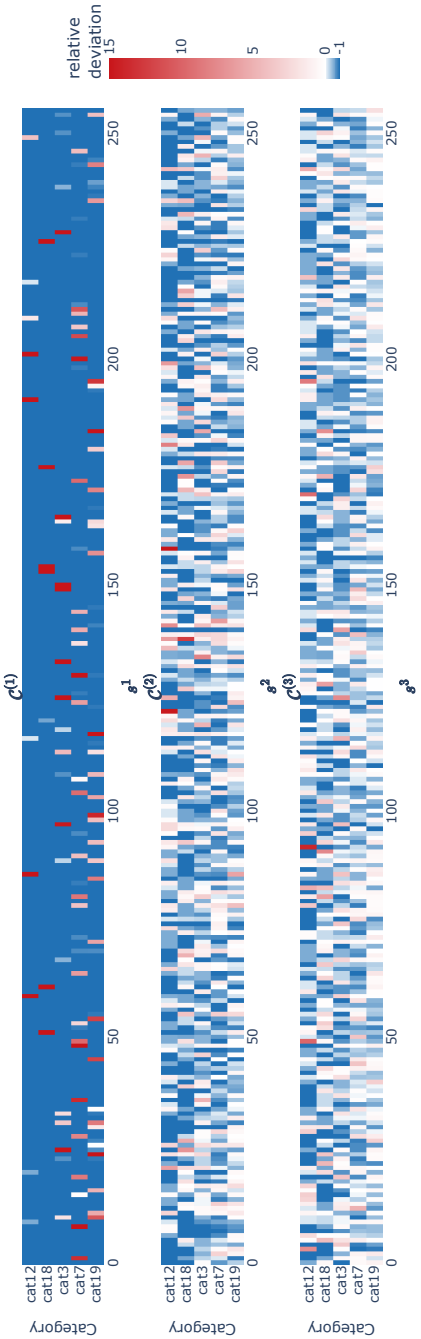


Figure 5.7: Codebook distribution heatmaps for $\mathcal{C}^{(1)}$, $\mathcal{C}^{(2)}$, and $\mathcal{C}^{(3)}$ over sampled categories on the Instacart dataset. A red band indicates that a category strongly prefers a specific semantic ID value, white represents a near-uniform distribution, while a blue region indicates underrepresentation.

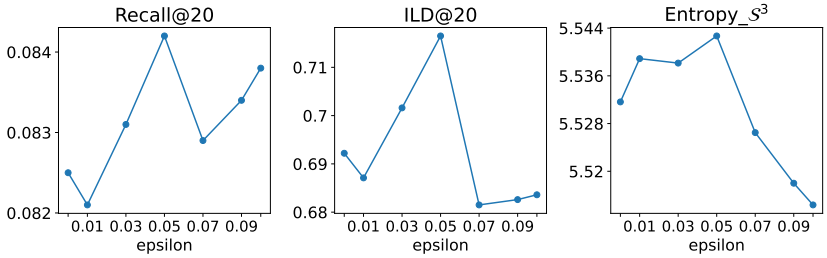


Figure 5.8: Varying ϵ configurations on the third layer of semantic IDs $\mathcal{S}^3$ for the Instacart dataset.

the results. We observe a clear co-movement among the three metrics: at $\epsilon = 0.05$, the codebook entropy reaches its maximum, and Recall@20 and ILD@20 simultaneously achieve their highest values. This indicates that increased codebook entropy is positively associated with both diversity gains and accuracy preservation.

We hypothesize that this empirical correlation reflects a layer-by-layer amplification effect along the following causal chain: (i) Higher codewords assignment entropy at $\mathcal{S}^3$ produces a more uniform semantic ID distribution across the item catalog. (ii) This uniformity propagates to the training data distribution, exposing the generative model to a broader and more balanced set of semantic IDs during training. (iii) As a result, the model learns a less peaked predictive distribution at inference time. (iv) A flatter predictive distribution expands the effective set of semantic IDs reachable by beam search, allowing the model to recommend a more diverse set of items. (v) Although semantic IDs do not strictly map to categories, the increased spread over semantic IDs translates into greater category-level diversity in the final recommendations.

A more general intuition is that altering the distribution of semantic IDs at the source inevitably propagates to the distribution of final recommendations, and this redistribution, when measured across categories, manifests as improved diversity. A rigorous verification of each link in this causal chain, and in particular the quantitative relationship between codebook entropy, predictive entropy, and downstream diversity, is left to future work.

5.6 Related Work

5.6.1 Next basket recommendation

Discriminative next basket recommendation. Next basket recommendation (NBR) aims to predict a set of items that match a user's preferences based on historical basket sequences. Existing methods model dependencies across baskets [76] and item correlations within each basket [61, 138]. While effective, they mainly focus on accuracy-oriented prediction and rely on discriminative scoring [44, 66]. Recent studies on *repetitive and exploratory* recommendation show that strong NBR methods often favor items that users have already purchased [63, 65, 77]. They also show that evaluating overall accuracy alone can hide the different behavior of repetitive and exploratory recommendations [65]. Therefore, recent work often studies repetitive [3,

54] and exploratory [61] NBR separately. Our work focuses on exploratory NBR.

Generative next basket recommendation. Generative NBR formulates next-basket prediction as a generation problem, where the model directly generates the item identifiers of the future basket. This paradigm is still under-explored. Existing studies include DiffNBR [147], which uses diffusion models to capture spatial and temporal purchasing patterns and disentangle repetitive and exploratory behaviors, and GenRec [108], which uses an autoregressive decoder to generate next-basket items one by one while modeling item correlations inside the target basket. These methods mainly focus on improving prediction accuracy or modeling basket behavior, while the diversity of generated baskets and the structure of the generation target space are less explicitly considered.

5.6.2 Generative recommendation with semantic IDs

Generative recommendation formulates item recommendation as a sequence generation task, where a model generates item identifiers conditioned on user interaction histories [51, 99]. Existing studies mainly explore three types of item identifiers: ID-based identifiers [42, 46], textual identifiers [6, 91], and codebook-based identifiers [74, 99, 120, 150]. Among them, codebook-based semantic IDs are widely used because they provide compact and semantically structured item representations. For example, TIGER trains a residual quantized VAE to convert item content embeddings into semantic IDs, where similar items tend to share identifier prefixes [99].

Prior work mainly uses Sinkhorn-Knopp to resolve semantic ID collisions, either during RQ-VAE training or as a post-processing step for redistributing conflicting items [59, 144, 150]. In contrast, our method applies Sinkhorn-Knopp to all items at selected quantization layers, aiming to improve the distribution of semantic IDs and enhance downstream recommendation diversity.

Diversity optimization. Generative recommendation often suffers from token-level bias amplification, producing highly homogenized lists. Prior work mitigates this at the decoding stage through techniques such as temperature-based sampling [99] and diverse beam search [91], but these approaches typically incur a strict accuracy-diversity trade-off [112]. In this work, we address this issue by constructing diversity-aware semantic IDs and using them as generation targets in a mask-prompted generative recommender.

5.7 Conclusion

In this chapter, we have studied exploratory NBR in the semantic-ID-based generative recommendation paradigm to address RQ4:

> *How can we design a semantic-ID-based generative recommendation method for exploratory basket recommendation, and how to improve recommendation diversity under the generative paradigm?*

We have proposed MPGBR, which combines diversity-aware semantic ID construction with mask-prompted basket generation. By injecting basket co-occurrence signals and balancing code assignments through optimal transport, MPGBR yields more entropic

semantic ID assignments, thereby improving downstream recommendation diversity. By reconstructing masked item IDs from basket context, the model captures item relations without imposing a fixed order over the whole basket. Experiments on three datasets show that MPGBR achieves competitive accuracy while improving diversity-oriented performance.

Implications. Our findings show that diversity in generative recommendation can be improved during semantic ID construction. By increasing the entropy of codebook assignments, the model reduces token-level bias before it is amplified by autoregressive generation. Although semantic IDs do not explicitly represent item categories, they form a discrete representation space that shapes downstream generation. We further argue that semantic IDs should be grounded in item content, while using CF signals as a complement. CF-only IDs are restricted to items observed during training, whereas content-grounded IDs can cover the full catalog and better generalize to new items.

Limitations and future work. First, our experiments focus on grocery datasets, where baskets naturally reflect co-purchase behavior. Future work could examine whether the framework generalizes to other basket-style domains, such as music playlists and fashion outfits. Second, MPGBR balances selected codebooks, but the optimal balancing granularity may vary across datasets. Future work could explore adaptive strategies for selecting codebooks and adjusting the balancing strength. Finally, we plan to study whether interventions at the semantic ID stage can support other distribution-related objectives, such as popularity debiasing and fairness.

5.8 Appendix

To facilitate reproducibility, we release the source code and all hyperparameters in an online repository: https://github.com/lynEcho/MPGBR

6 Conclusions

In this thesis, we have studied the evaluation and optimization of beyond-accuracy objectives for next basket recommendation (NBR), covering item fairness, diversity, and repeat bias. We have investigated the balance between repeat and explore items obtained from NBR methods and its impact on the joint evaluation protocol of accuracy and beyond-accuracy objectives. Moving from discriminative to semantic-ID-based generative paradigms, we have explored how beyond-accuracy objectives perform and how bias arises under this emerging mechanism.

6.1 Summary of Findings

RQ1 *What is the item fairness performance of current next basket recommendation methods?*

This question has been answered in Chapter 2, where we systematically evaluate the item fairness of representative NBR methods on three grocery shopping datasets. Specifically, we apply seven merit-based and distribution-based fair ranking metrics to assess exposure fairness between popular and unpopular item groups. We further examine the robustness of item fairness metrics with respect to basket size, position weighting models, and user repeat behavior.

Through our experiments, we arrive at several important findings: (i) the fairness metrics considered, including equal opportunity and statistical parity, rank the NBR methods differently. Nevertheless, most metrics agree that repeat-biased methods are fairer than explore-biased ones. (ii) Contrary to the observations in [98], inequality of amortized attention (IAA) emerges as the metric most sensitive to changes in both basket size and position weighting models. (iii) Across user groups exhibiting different degrees of repeat purchase behavior, the fairness metrics reveal clear differences, indicating that repeat purchase behavior does influence item fairness, though the patterns are not consistent across metrics. (iv) Taken together, our results suggest that expected exposure loss (EEL) and expected exposure disparity (EED) are the most suitable metrics for NBR: they are highly robust to different parameter configurations and different position weighting models, and they support fairness measurements over multiple groups.

Our work confirms that fair ranking metrics exhibit different sensitivity patterns across information access systems, owing to the characteristics of each scenario. For

rigor, we recommend selecting fair ranking metrics according to the specific task and the reliability of the relevance labels, and testing the sensitivity of these metrics for each scenario, following the evaluation framework from [98], to ensure that the chosen metrics are reliable. This chapter is the first attempt to study item exposure fairness in the NBR domain, providing a reference point for the optimization of future NBR methods. We have focused primarily on group fairness defined by popularity, but other grouping ways, such as by brand or category, are also worth examining.

RQ2 *How can we design a basket recommendation method that jointly targets accuracy and beyond-accuracy objectives, and does the current evaluation paradigm adequately support this goal?*

This question has been answered in Chapter 3, where we propose a short-cut strategy that trades off accuracy for exploration and uses explore items to optimize beyond-accuracy metrics for NBR, and where we design a two-step repetition-exploration (TREx) framework to verify the effectiveness of this strategy. We evaluate TREx on two datasets across eight beyond-accuracy metrics. The results show that: (i) exploiting "short-cuts" via TREx yields better performance on a wide range of metrics. Moreover, we observe that the more strongly a beyond-accuracy metric correlates with accuracy, the harder it is to improve that metric through the short-cut strategy. (ii) However, this reveals a loophole in the current overall evaluation paradigm: by evaluating a basket as a whole, we mask the poor accuracy of exploratory recommendations. We therefore propose a fine-grained evaluation paradigm that assesses repetitive and exploratory recommendations separately, serving as a supplement to evaluate NBR methods more comprehensively.

This chapter sheds light on a research direction that considers both accuracy and beyond-accuracy metrics in NBR. We refine the evaluation protocol, offering a reference for follow-up research that should recognize the short-cut and the pitfalls in original evaluation paradigm.

RQ3 *How can we design algorithms to jointly improve diversity and item fairness while mitigating repeat bias in next basket recommendation?*

This question has been answered in Chapter 4, where we propose mixed-integer linear programming (MILP)-based re-ranking methods to mitigate the heavy repeat bias of current NBR methods while optimizing diversity or item fairness. The proposed repeat-bias-aware diversity optimization (RADiv) and repeat-bias-aware item fairness optimization (RAIF) can be adapted to both the unified and combined NBR paradigms. Through experiments on three real-world retail datasets, we find that: (i) the proposed RADiv and RAIF algorithms effectively enhance diversity and item fairness while mitigating repeat bias, incurring only an acceptable utility loss. (ii) In NBR, a trade-off persists between accuracy and beyond-accuracy objectives such as item fairness, diversity, and repeat bias.

In many cases, NBR methods can achieve higher Recall by recommending more repeat items (i.e., heavier repeat bias), but this reduces recommendation novelty and serendipity, hurting long-term user satisfaction. In this thesis, we assume that the ideal distribution between repeat and explore items aligns with the mean of ground truth

baskets; however, the definition and quantification of repeat bias remain open questions. Since repeat bias is therefore an unavoidable issue when evaluating NBR methods, one cleaner yet effective option is to split NBR into repeat-only [3, 54] and explore-only [61] tasks.

RQ4 *How can we design a semantic-ID-based generative recommendation method for exploratory basket recommendation, and how to improve recommendation diversity under the generative paradigm?*

This question has been answered in Chapter 5, where we shift to a semantic-ID-based generative paradigm and focus on exploratory recommendation in NBR. Specifically, we propose a mask-prompted generative framework that uses a bidirectional transformer to encode historical baskets and a lightweight decoder to reconstruct the semantic IDs of masked items. Moreover, we propose a balanced item tokenizer that aims to improve downstream recommendation diversity by increasing the codeword assignment entropy.

We conduct experiments on three datasets and find that: (i) mask-prompted generative basket recommendation (MPGBR) achieves competitive accuracy while remaining on the accuracy-diversity Pareto frontier. (ii) Balanced RQ-KMeans increases codeword assignment entropy through optimal transport, thereby improving diversity at the representation stage without compromising accuracy. (iii) We further argue that semantic IDs should be grounded in item metadata and complemented, rather than defined, by collaborative filtering (CF) signals, because content-grounded semantic IDs can cover the full catalog and generalize to new items.

Although our experiments empirically confirm that increasing the entropy of codeword assignments improves downstream recommendation diversity, the underlying mechanism remains unclear. We hypothesize that token-level bias is reduced at the semantic ID construction stage before it is amplified by the generative model. However, verifying this mechanism, that is, measuring how bias propagates and how balanced tokens connect to diverse categories, remains challenging. Clarifying it could guide the design of optimization algorithms for other distribution-based objectives, such as popularity bias and exposure fairness.

6.2 Future Work

Each chapter in this thesis lists chapter-specific limitations and directions for future work in its conclusion section. Here, we step further to discuss broader avenues for trustworthy recommender systems under evolving paradigms.

6.2.1 Discriminative vs. semantic-ID-based generative paradigms

Recommender systems are undergoing a paradigm shift from discriminative to generative modeling, driven by several appealing properties of the generative formulation. First, generative architectures exhibit favorable scaling behavior: HSTU [143] demonstrates that recommendation quality continues to improve with model and compute scale, echoing the scaling laws observed in large language models. Second, the conventional multi-stage cascaded pipeline (retrieval, ranking, re-ranking) suffers from objective

inconsistency across stages and cascaded error propagation. The generative paradigm addresses this by replacing the pipeline with a single end-to-end model that unifies retrieval and ranking [27, 152]. Third, semantic IDs offer a principled solution to item representation. Instead of maintaining a massive embedding table with one independent entry per item, they encode all items as short token sequences over a compact shared vocabulary, so that semantically similar items share prefix tokens. This enables natural generalization to cold-start items and reduces memory footprint by orders of magnitude [99].

However, whether these theoretical advantages translate into consistent empirical gains, across both accuracy and beyond-accuracy objectives, remains an open question. Lepage et al. [59] show that semantic-ID-based generative recommendation and a well-tuned SASRec achieve comparable accuracy on the Amazon 2014 and 2023 datasets. Yang et al. [132] reveal that although larger models improve accuracy, they further skew recommendations toward popular items. All this evidence prompts us to compare and rethink these two paradigms through both mechanistic and empirical analysis.

Mechanistic comparison. Prior work in information retrieval has analyzed the mechanisms underlying generative and dense retrieval [89]. For example, under some assumptions, generative retrieval has been shown to be a special case of multi-vector dense retrieval, with both paradigms computing query–document relevance under a common formulation [124]. However, whether this equivalence transfers to generative recommendation remains unexamined. Taking discriminative recommenders (e.g., SASRec) as the natural reference point, generative recommenders (e.g., TIGER) differ primarily along three axes: (i) *item representation*, an independent embedding table vs. compositional semantic codes; (ii) *probability normalization*, locally via negative sampling when the item catalog outgrows exact softmax vs. globally over a compact token vocabulary; and (iii) *inference*, nearest-neighbor lookup vs. autoregressive generation.

A mechanistic comparison between discriminative and generative recommenders is valuable for several reasons. (i) It supports attribution: only by decomposing the two paradigms into their differing components (e.g., item representation, normalization, and inference) can we credit empirical performance gaps to specific mechanisms rather than to "the paradigm" as a whole. (ii) It tells us when each paradigm should prevail: once we understand how the mechanisms differ, we can predict which paradigm performs better under specific conditions (e.g., different dataset scales or long-tail settings) and verify that prediction empirically. (iii) It is a prerequisite for principled beyond-accuracy optimization. Since interventions on unfairness or popularity bias must target the stage where the problem arises, e.g., quantization, training, or decoding, we need mechanistic analyses to trace an observed bias back to its source.

6.2.2 Fairness in recommendation

Although fairness has been extensively studied in discriminative recommendation [31, 142], newer paradigms such as LLM-based and semantic-ID-based generative recommendation change how bias is injected and amplified [50, 73], so fairness problems resurface in new forms. We organize fairness research in recommender systems as a four-stage loop: (i) recognizing unfairness on the user or provider/item side; (ii) measuring it with metrics grounded in different normative principles [9, 68, 104]; (iii)

mitigating it through algorithms and interventions [86]; and (iv) monitoring its long-term impact in the feedback loop [2, 126]. In the semantic-ID-based generative paradigm, however, each step of this loop faces new challenges, which are further entangled with problems left unresolved from the discriminative era.

Recognize. The unfairness identified in the discriminative paradigm (e.g., skewed exposure allocation across items and utility disparities across user groups) persists, because its root cause, biased interaction data, remains unchanged. Moreover, the generative paradigm widens the space where unfairness can enter by introducing new sources of bias. Bias can now be injected at tokenization and amplified during decoding [73]. This may give rise to qualitatively new forms of unfairness, such as exposure disparities driven by prefix assignment rather than merit.

Measure. Fairness metrics themselves largely carry over across paradigms, since the output remains a ranked list. What is unsettled are the normative choices hidden inside their definitions, and in recommendation these choices remain contested. First, what are comparable item groups? Items are typically grouped by category or popularity [50], but it is unclear that such groupings define a reasonable basis for exposure comparison. Second, what should stand in for merit? Common proxies are predicted relevance or ground-truth labels [98], yet both can inherit biases. Third, on what grounds should users be grouped? Sensitive attributes or activity levels are often used [68], but requiring equal recommendation accuracy across users with different activity levels may set an unreasonable bar, as inactive users provide far less interaction data. Making these choices explicit and justifying them within each platform's context is what turns a metric from an abstract quantity into a defensible fairness objective. This step enables the measurement to translate into real-world impact rather than remaining conceptual.

Mitigate. In discriminative recommendation, fairness interventions have been organized according to the stages of the machine learning pipeline: pre-processing the data, regularizing the training objective, and re-ranking the output [127]. The generative paradigm preserves this logic but redefines the stages. Early efforts have emerged at each stage: regularizing semantic tokens [34], reshaping the probability allocation in training [73, 132, 135], and rebalancing during decoding. However, these interventions target only one or two stages at a time, without quantifying each stage's contribution to the bias accumulated across the pipeline. We therefore call for a systematic cross-stage comparison, examining how interventions at each stage trade fairness against accuracy. This way, the choice of intervention stage is driven by diagnosis, and its trade-offs are owned explicitly.

Monitor. Even within the discriminative paradigm, developing long-term fairness algorithms and monitoring their actual impact across feedback loops remains the least developed step. Auditing such long-horizon dynamics requires simulation, but the simulation itself also brings some unresolved issues. First, how to model the co-evolution of multiple stakeholders: users adapt their preferences and behavior to what they are shown [115], while providers adjust their offerings in response to the engagement they receive [134]. Second, LLM-powered agents are a promising route to simulating stakeholders' long-term behavior, but they carry biases of their own that must be accounted for [145]. The semantic-ID-based generative paradigm adds a new link to the feedback loop. Today's recommendations produce tomorrow's interaction data, which retrains not

only the recommender but also the tokenizer. Semantic IDs therefore drift as new items and interactions arrive [35], and bias can accumulate as the codebook itself evolves.

We organize fairness research as a loop of four steps, viz. recognize, measure, mitigate, and monitor, and apply it to the shift from the discriminative to the semantic-ID-based generative paradigm. Fairness concerns arise whenever a technology enters the real world and begins allocating benefits and harms among people. As recommendation evolves from matrix factorization toward possibly agent-based systems, the fairness loop remains. The accumulated experience and frameworks of fairness research therefore transfer to each new paradigm, requiring only adaptation to its mechanisms. Fairness, as a dimension of trustworthiness, is ultimately a value choice rather than a technical inevitability. It is this choice that motivates us to build recommender systems that are fair, accountable, and responsible. This perspective grows only more urgent as people delegate ever more judgment, decisions, and action to algorithms.

Bibliography

[1] H. Abdollahpouri, M. Mansoury, R. Burke, and B. Mobasher. The unfairness of popularity bias in recommendation. In *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems*, 2019. (Cited on page 14.)

[2] N.-J. Akpinar, C. DiCiccio, P. Nandy, and K. Basu. Long-term dynamics of fairness intervention in connection recommender systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 22–35, 2022. (Cited on page 93.)

[3] M. Ariannezhad, S. Jullien, M. Li, M. Fang, S. Schelter, and M. de Rijke. ReCANet: A repeat consumption-aware neural network for next basket recommendation in grocery shopping. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1240–1250, 2022. (Cited on pages 2, 13, 15, 16, 30, 33, 39, 40, 56, 57, 67, 74, 85, and 91.)

[4] M. Ariannezhad, M. Li, S. Jullien, and M. de Rijke. Complex item set recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3444–3447, 2023. (Cited on pages 12, 29, and 50.)

[5] T. Bai, J.-Y. Nie, W. X. Zhao, Y. Zhu, P. Du, and J.-R. Wen. An attribute-aware neural attentive model for next basket recommendation. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1201–1204, 2018. (Cited on page 33.)

[6] K. Bao, J. Zhang, W. Wang, Y. Zhang, Z. Yang, Y. Luo, C. Chen, F. Feng, and Q. Tian. A bi-step grounding paradigm for large language models in recommendation systems. *ACM Transactions on Recommender Systems*, 3(4):1–27, 2025. (Cited on page 86.)

[7] C. Bauer, C. Bagchi, O. A. Hundogan, and K. van Es. Where are the values? A systematic literature review on news recommender systems. *ACM Transactions on Recommender Systems*, 2(3):23:1–23:40, 2024. (Cited on page 51.)

[8] S. Bi, C. Wang, W. Wang, Y. Mou, F. Feng, T. Biao, and P. Yan. ReList: A multi-objective reasoning framework for diversified listwise query recommendation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pages 1392–1405, 2026. (Cited on page 68.)

[9] A. J. Biega, K. P. Gummadi, and G. Weikum. Equity of attention: Amortizing individual fairness in rankings. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 405–414, 2018. (Cited on pages 1, 17, 18, 35, 52, and 92.)

[10] J. Bobadilla, R. Lara-Cabrera, Á. González-Prieto, and F. Ortega. DeepFair: Deep learning for improving fairness in recommender systems. *arXiv preprint arXiv:2006.05255*, 2020. (Cited on page 33.)

[11] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, 1998. (Cited on pages 1, 2, and 51.)

[12] Y. Cen, J. Zhang, X. Zou, C. Zhou, H. Yang, and J. Tang. Controllable multi-interest framework for recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2942–2951, 2020. (Cited on page 36.)

[13] A. J. Chaney, B. M. Stewart, and B. E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 224–232, 2018. (Cited on page 1.)

[14] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 621–630, 2009. (Cited on page 51.)

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

[16] L. Chen, G. Zhang, and E. Zhou. Fast greedy MAP inference for determinantal point process to improve recommendation diversity. *Advances in Neural Information Processing Systems*, 31:5627–5638, 2018. (Cited on page 2.)

[17] M. Chen. Exploration in recommender systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 551–553, 2021. (Cited on page 4.)

[18] M. Chen, Y. Wang, C. Xu, Y. Le, M. Sharma, L. Richardson, S.-L. Wu, and E. Chi. Values of user

exploration in recommender systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 85–95, 2021. (Cited on page 67.)

[19] W. Chen, P. Ren, F. Cai, F. Sun, and M. de Rijke. Improving end-to-end sequential recommendations with intent-aware diversification. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 175–184, 2020. (Cited on pages 33, 36, and 46.)

[20] W. Chen, P. Ren, F. Cai, F. Sun, and M. de Rijke. Multi-interest diversification for end-to-end sequential recommendation. *ACM Transactions on Information Systems*, 40(1):1–30, 2021. (Cited on page 30.)

[21] Y. Chen, J. Li, C. Liu, C. Li, M. Anderle, J. McAuley, and C. Xiong. Modeling dynamic attributes for next basket recommendation. *arXiv preprint arXiv:2109.11654*, 2021. (Cited on page 30.)

[22] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the International Conference on Web Search and Web Data Mining*, pages 87–94, 2008. (Cited on page 12.)

[23] P. Cremonesi, Y. Koren, and R. Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the 4th ACM Conference on Recommender Systems*, pages 39–46, 2010. (Cited on page 1.)

[24] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26:2292–2300, 2013. (Cited on page 72.)

[25] M. de Rijke. Beyond accuracy goals, again. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining*, pages 2–3, 2023. (Cited on page 51.)

[26] Y. Deldjoo, Z. He, J. McAuley, A. Korikov, S. Sanner, A. Ramisa, R. Vidal, M. Sathiamoorthy, A. Kasirzadeh, and S. Milano. A review of modern recommender systems using generative models (Gen-RecSys). In *Proceedings of the 30th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 6448–6458, 2024. (Cited on page 1.)

[27] J. Deng, S. Wang, K. Cai, L. Ren, Q. Hu, W. Ding, Q. Luo, and G. Zhou. OneRec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *arXiv preprint arXiv:2502.18965*, 2025. (Cited on pages 71, 72, and 92.)

[28] F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, and B. Carterette. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 275–284, 2020. (Cited on pages 14, 17, 18, and 35.)

[29] Dunnhumby. Source files, 2024. `https://www.dunnhumby.com/source-files/`. (Cited on pages 56 and 74.)

[30] M. D. Ekstrand, R. Burke, and F. Diaz. Fairness and discrimination in retrieval and recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1403–1404, 2019. (Cited on pages 30 and 33.)

[31] M. D. Ekstrand, A. Das, R. Burke, and F. Diaz. Fairness in information access systems. *Foundations and Trends in Information Retrieval*, 16(1-2):1–177, 2022. (Cited on pages 1, 11, 14, and 92.)

[32] M. D. Ekstrand, B. Carterette, and F. Diaz. Distributionally-informed recommender system evaluation. *ACM Transactions on Recommender Systems*, 2(1):6:1–6:27, 2023. (Cited on page 12.)

[33] G. Faggioli, M. Polato, and F. Aiolli. Recency aware collaborative filtering for next basket recommendation. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 80–87, 2020. (Cited on pages 2, 13, 15, 16, 30, 33, 37, 39, 40, 49, 56, 57, 74, and 75.)

[34] Z. Fan, Z. Chen, L. Ma, J. Huang, L. Morishetti, K. Nag, S. Kumar, and K. Achan. CRAB: Codebook rebalancing for bias mitigation in generative recommendation. *arXiv preprint arXiv:2604.05113*, 2026. (Cited on pages 2 and 93.)

[35] Y. Feng, J. Liu, M. Han, D. Li, H. Gu, P. Zhang, T. Lu, and N. Gu. Drift-aware continual tokenization for generative recommendation. *arXiv preprint arXiv:2603.29705*, 2026. (Cited on page 94.)

[36] S. Flaxman, S. Goel, and J. M. Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80:298–320, 2016. (Cited on page 61.)

[37] J. Fu, X. Ge, A. Karatzoglou, I. Arapakis, S. Verberne, J. M. Jose, and Z. Ren. Differentiable semantic ID for generative recommendation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 369–379, 2026. (Cited on page 68.)

[38] I. Gabriel, A. Manzini, G. Keeling, L. A. Hendricks, V. Rieser, H. Iqbal, N. Tomašev, I. Ktena, Z. Kenton, M. Rodriguez, et al. The ethics of advanced AI assistants. *arXiv preprint arXiv:2404.16244*, 2024. (Cited on page 2.)

[39] Y. Ge, S. Zhao, H. Zhou, C. Pei, F. Sun, W. Ou, and Y. Zhang. 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, 2020. (Cited on page 61.)

[40] Y. Ge, S. Liu, R. Gao, Y. Xian, Y. Li, X. Zhao, C. Pei, F. Sun, J. Ge, W. Ou, et al. Towards long-term

fairness in recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 445–453, 2021. (Cited on pages 18 and 30.)

[41] Y. Ge, S. Liu, Z. Fu, J. Tan, Z. Li, S. Xu, Y. Li, Y. Xian, and Y. Zhang. A survey on trustworthy recommender systems. *ACM Transactions on Recommender Systems*, 3(2):1–68, 2024. (Cited on page 1.)

[42] S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang. Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 299–315, 2022. (Cited on pages 1 and 86.)

[43] M. Goyani and N. Chaurasiya. A review of movie recommendation system: Limitations, survey and challenges. *Electronic Letters on Computer Vision and Image Analysis*, 19(3):18–37, 2020. (Cited on page 13.)

[44] H. Hu and X. He. Sets2Sets: Learning from sequential sets with neural networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1491–1499, 2019. (Cited on pages 2, 15, 30, 33, and 85.)

[45] H. Hu, X. He, J. Gao, and Z.-L. Zhang. Modeling personalized item frequency information for next-basket recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1071–1080, 2020. (Cited on pages 2, 13, 16, 18, 30, 33, 37, 40, 49, 56, 57, and 75.)

[46] W. Hua, S. Xu, Y. Ge, and Y. Zhang. How to index item IDs for recommendation foundation models. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 195–204, 2023. (Cited on page 86.)

[47] J. Hutter, H. C. Bakker, S. Fris, M. Bernardy, and Y. Liu. A systematic reproducibility study of BSARec for sequential recommendation. In *Proceedings of the 48th European Conference on Information Retrieval*, pages 17–32, 2026.

[48] Instacart. Market basket analysis, 2017. `https://www.kaggle.com/datasets/yasserh/instacart-online-grocery-basket-analysis-dataset`. (Cited on pages 56 and 74.)

[49] D. Jannach. Multi-objective recommender systems: Survey and challenges. In *Proceedings of the 2nd Workshop on Multi-Objective Recommender Systems held in conjunction with the 16th ACM Conference on Recommender Systems*, 2022. (Cited on page 30.)

[50] M. Jiang, K. Bao, J. Zhang, W. Wang, Z. Yang, F. Feng, and X. He. Item-side fairness of large language model-based recommendation system. In *Proceedings of the ACM Web Conference 2024*, pages 4717–4726, 2024. (Cited on pages 92 and 93.)

[51] C. M. Ju, L. Collins, L. Neves, B. Kumar, L. Y. Wang, T. Zhao, and N. Shah. Generative recommendation with semantic IDs: A practitioner's handbook. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6420–6425, 2025. (Cited on page 86.)

[52] M. Kaminskas and D. Bridge. Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems*, 7(1):1–42, 2016. (Cited on pages 2 and 51.)

[53] W.-C. Kang and J. McAuley. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining*, pages 197–206, 2018. (Cited on page 1.)

[54] O. Katz, O. Barkan, N. Koenigstein, and N. Zabari. Learning to ride a buy-cycle: A hyper-convolutional model for next basket repurchase recommendation. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 316–326, 2022. (Cited on pages 2, 13, 18, 30, 33, 56, 86, and 91.)

[55] D. Kowald, M. Schedl, and E. Lex. The unfairness of popularity bias in music recommendation: A reproducibility study. In *Proceedings of the 42nd European Conference on Information Retrieval*, pages 35–42, 2020. (Cited on page 14.)

[56] A. Kumar and K. Hosanagar. Measuring the value of recommendation links on product demand. *Information Systems Research*, 30(3):819–838, 2019. (Cited on page 29.)

[57] D. Le, H. W. Lauw, and Y. Fang. Correlation-sensitive next-basket recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2808–2814, 2019. (Cited on pages 2, 13, 30, and 33.)

[58] Y. Leng, L. Yu, J. Xiong, and G. Xu. Recurrent convolution basket map for diversity next-basket recommendation. In *Proceedings of the 25th International Conference on Database Systems for Advanced Applications*, pages 638–653, 2020. (Cited on pages 33, 46, 50, and 51.)

[59] S. Lepage, J. Mary, and D. Picard. Closing the performance gap in generative recommenders with collaborative tokenization and efficient modeling. *arXiv preprint arXiv:2508.14910*, 2025. (Cited on pages 71, 72, 73, 86, and 92.)

[60] M. Li, M. Ariannezhad, A. Yates, and M. de Rijke. Who will purchase this item next? Reverse next

period recommendation in grocery shopping. *ACM Transactions on Recommender Systems*, 1(2):1–32, 2023. (Cited on pages 13, 30, 47, and 50.)

[61] M. Li, M. Ariannezhad, A. Yates, and M. de Rijke. Masked and swapped sequence modeling for next novel basket recommendation in grocery shopping. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 35–46, 2023. (Cited on pages 2, 13, 30, 33, 56, 67, 73, 74, 75, 85, 86, and 91.)

[62] M. Li, J. Huang, and M. de Rijke. Repetition and exploration in offline reinforcement learning-based recommendations. In *Proceedings of the Workshop on Deep Reinforcement Learning for Information Retrieval at CIKM'23*, 2023. (Cited on pages 30, 47, and 50.)

[63] M. Li, S. Jullien, M. Ariannezhad, and M. de Rijke. A next basket recommendation reality check. *ACM Transactions on Information Systems*, 41(4):1–29, 2023. (Cited on pages 2, 13, 14, 15, 16, 18, 22, 23, 29, 30, 33, 34, 38, 39, 40, 46, 47, 51, 56, 57, 67, 74, and 85.)

[64] M. Li, A. Vardasbi, A. Yates, and M. de Rijke. Repetition and exploration in sequential recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2532–2541, 2023. (Cited on pages 30, 47, and 50.)

[65] M. Li, Y. Liu, S. Jullien, M. Ariannezhad, A. Yates, M. Aliannejadi, and M. de Rijke. Are we really achieving better beyond-accuracy performance in next basket recommendation? In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 924–934, 2024. (Cited on pages 4, 50, 51, 55, 56, 57, 60, 67, 74, and 85.)

[66] R. Li, L. Zhang, G. Liu, and J. Wu. Next basket recommendation with intent-aware hypergraph adversarial network. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1303–1312, 2023. (Cited on pages 2, 49, and 85.)

[67] X. Li, Z. Liu, L. Ma, K. Nag, S. Guo, P. S. Yu, and K. Achan. Mitigating frequency bias in next-basket recommendation via deconfounders. In *2022 IEEE International Conference on Big Data*, pages 616–625, 2022. (Cited on page 14.)

[68] Y. Li, H. Chen, Z. Fu, Y. Ge, and Y. Zhang. User-oriented fairness in recommendation. In *Proceedings of the Web Conference 2021*, pages 624–632, 2021. (Cited on pages 1, 14, 51, 52, 92, and 93.)

[69] Y. Li, H. Chen, S. Xu, Y. Ge, J. Tan, S. Liu, and Y. Zhang. Fairness in recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 14(5):95:1–95:48, 2023. (Cited on pages 11 and 40.)

[70] Y. Li, X. Lin, W. Wang, F. Feng, L. Pang, W. Li, L. Nie, X. He, and T.-S. Chua. A survey of generative search and recommendation in the era of large language models. *arXiv preprint arXiv:2404.16924*, 2024. (Cited on page 1.)

[71] Y. Liang, T. Qian, Q. Li, and H. Yin. Enhancing domain-level and user-level adaptivity in diversified recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 747–756, 2021. (Cited on pages 36 and 53.)

[72] X. Lin, H. Chen, C. Pei, F. Sun, X. Xiao, H. Sun, Y. Zhang, W. Ou, and P. Jiang. A pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 20–28, 2019. (Cited on page 2.)

[73] X. Lin, P. Liu, W. Wang, Y. Hu, C. Xu, F. Feng, Q. Wang, and T.-S. Chua. Bringing reasoning to generative recommendation through the lens of cascaded ranking. In *Proceedings of the ACM Web Conference 2026*, pages 8939–8949, 2026. (Cited on pages 2, 92, and 93.)

[74] E. Liu, B. Zheng, C. Ling, L. Hu, H. Li, and W. X. Zhao. Generative recommender with end-to-end learnable item tokenization. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 729–739, 2025. (Cited on pages 68 and 86.)

[75] Y. Liu, M. Li, M. Ariannezhad, M. Mansoury, M. Aliannejadi, and M. de Rijke. Measuring item fairness in next basket recommendation: A reproducibility study. In *Proceedings of the 46th European Conference on Information Retrieval*, pages 210–225, 2024. (Cited on pages 31, 33, 34, 35, 39, 40, 50, 51, 56, and 74.)

[76] Y. Liu, M. Liu, C. Walder, and L. Xie. A universal sets-level optimization framework for next set recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1544–1554, 2024. (Cited on page 85.)

[77] Y. Liu, M. Li, M. Aliannejadi, and M. de Rijke. Repeat-bias-aware optimization of beyond-accuracy metrics for next basket recommendation. In *Proceedings of the 47th European Conference on Information Retrieval*, pages 214–229, 2025. (Cited on pages 68 and 85.)

[78] Y. Liu, Y. Tang, M. Li, C. Xu, M. Aliannejadi, and M. de Rijke. Mask-prompted generative retrieval for diverse and exploratory next basket recommendation. Under review, 2026.

[79] M. Ludewig and D. Jannach. Evaluation of session-based recommendation algorithms. *User Modeling*

and User-Adapted Interaction, 28:331–390, 2018. (Cited on page 33.)

[80] Y. Ma, Y. He, X. Wang, Y. Wei, X. Du, Y. Fu, and T.-S. Chua. MultiCBR: Multi-view contrastive learning for bundle recommendation. *ACM Transactions on Information Systems*, 42(4):1–23, 2024. (Cited on page 1.)

[81] M. Mansoury, H. Abdollahpouri, M. Pechenizkiy, B. Mobasher, and R. Burke. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 2145–2148, 2020. (Cited on page 1.)

[82] S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *Extended Abstracts Proceedings of the 2006 Conference on Human Factors in Computing Systems*, pages 1097–1101, 2006. (Cited on page 1.)

[83] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):1–27, 2008. (Cited on pages 18 and 51.)

[84] L. A. Møller. Between personal and public interest: How algorithmic news recommendation reconciles with journalism as an ideology. *Digital Journalism*, 10(10):1794–1812, 2022. (Cited on page 1.)

[85] M. Morik, A. Singh, J. Hong, and T. Joachims. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 429–438, 2020. (Cited on page 33.)

[86] M. Naghiaei, H. A. Rahmani, and Y. Deldjoo. CPFair: Personalized consumer and producer fairness re-ranking for recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 770–779, 2022. (Cited on pages 2, 33, 52, 53, 57, and 93.)

[87] S. Naumov, M. Ananyeva, O. Lashinin, S. Kolesnikov, and D. I. Ignatov. Time-dependent next-basket recommendations. In *Proceedings of the 45th European Conference on Information Retrieval*, pages 502–511, 2023. (Cited on pages 13, 15, 39, 49, 56, 57, and 74.)

[88] H. Nguyen, Y. Liu, M. Mansoury, M. Aliannejadi, A. Hanjalic, and M. de Rijke. A reproducibility study of product-side fairness in bundle recommendation. In *Proceedings of the 19th ACM Conference on Recommender Systems*, pages 696–705, 2025.

[89] T. Nguyen and A. Yates. Generative retrieval as dense retrieval. *arXiv preprint arXiv:2306.11397*, 2023. (Cited on page 92.)

[90] OECD. E-commerce in the time of COVID-19. `https://doi.org/10.1787/3a2b78e8-en`, 2020. (Cited on page 29.)

[91] E. Palumbo, G. Penha, A. Damianou, J. L. R. García, T. C. Heath, A. Wang, H. Bouchard, and M. Lalmas. Text2Tracks: Prompt-based music recommendation via generative retrieval. *arXiv preprint arXiv:2503.24193*, 2025. (Cited on page 86.)

[92] Y.-J. Park and A. Tuzhilin. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, pages 11–18, 2008. (Cited on page 1.)

[93] G. K. Patro, A. Biswas, N. Ganguly, K. P. Gummadi, and A. Chakraborty. FairRec: Two-sided fairness for personalized recommendations in two-sided platforms. In *Proceedings of the Web Conference 2020*, pages 1194–1204, 2020. (Cited on page 1.)

[94] G. Penha, E. D'Amico, M. De Nadai, E. Palumbo, A. Tamborrino, A. Vardasbi, M. Lefarov, S. Lin, T. Heath, F. Fabbri, et al. Semantic IDs for joint generative search and recommendation. In *Proceedings of the 19th ACM Conference on Recommender Systems*, pages 1296–1301, 2025. (Cited on page 1.)

[95] A. Petrov and C. Macdonald. Aligning GPTRec with beyond-accuracy goals with reinforcement learning. In *Proceedings of the 2nd Workshop on Recommendation with Generative Models, in conjunction with the Web Conference 2024*, 2024. (Cited on page 2.)

[96] Y. Qin, P. Wang, and C. Li. The world is binary: Contrastive learning for denoising next basket recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 859–868, 2021. (Cited on pages 15 and 30.)

[97] M. Quadrana, P. Cremonesi, and D. Jannach. Sequence-aware recommender systems. *ACM Computing Surveys*, 51(4):1–36, 2018. (Cited on page 33.)

[98] A. Raj and M. D. Ekstrand. Measuring fairness in ranked results: An analytical and empirical comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 726–736, 2022. (Cited on pages 1, 3, 11, 12, 13, 14, 15, 16, 17, 18, 19, 23, 31, 34, 35, 36, 51, 57, 89, 90, and 93.)

[99] S. Rajput, N. Mehta, A. Singh, R. Hulikal Keshavan, T. Vu, L. Heldt, L. Hong, Y. Tay, V. Tran, J. Samost, et al. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315, 2023. (Cited on pages 2, 4, 68, 75, 86, and 92.)

[100] T. V. Rampisela, T. Ruotsalo, M. Maistro, and C. Lioma. Can we trust recommender system fairness evaluation? The role of fairness and relevance. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 271–281, 2024. (Cited on page 1.)

[101] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme. Factorizing personalized Markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 811–820, 2010. (Cited on pages 30 and 33.)

[102] C. Rus, J. Kareem, C. Xu, Y. Liu, Z. Deng, and M. Heuss. AMS42 at the NTCIR-18 FairWeb-2 task. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, pages 61–67, 2025.

[103] P. Sapiezynski, W. Zeng, R. E Robertson, A. Mislove, and C. Wilson. Quantifying the impact of user attention fair group representation in ranked lists. In *Companion Proceedings of the 2019 World Wide Web Conference*, pages 553–562, 2019. (Cited on pages 12 and 14.)

[104] A. Singh and T. Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2219–2228, 2018. (Cited on pages 1, 17, 35, 52, 53, and 92.)

[105] Y. Su, X. Wang, E. Y. Le, L. Liu, Y. Li, H. Lu, B. Lipshitz, S. Badam, L. Heldt, S. Bi, et al. Long-term value of exploration: Measurements, findings and algorithms. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 636–644, 2024. (Cited on pages 1, 50, and 67.)

[106] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang. 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, 2019. (Cited on pages 1 and 75.)

[107] L. Sun, Y. Bai, B. Du, C. Liu, H. Xiong, and W. Lv. Dual sequential network for temporal sets prediction. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1439–1448, 2020. (Cited on pages 30 and 33.)

[108] W. Sun, R. Xie, J. Zhang, W. X. Zhao, L. Lin, and J.-R. Wen. Generative next-basket recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 737–743, 2023. (Cited on pages 18, 50, 51, 68, 75, and 86.)

[109] TaFeng. Grocery dataset, 2001. `https://www.kaggle.com/datasets/chiranjivdas09/ta-feng-grocery-dataset`. (Cited on pages 56 and 74.)

[110] V.-A. Tran, G. Salha-Galvan, B. Sguerra, and R. Hennequin. Transformers meet ACT-R: Repeat-aware and sequential listening session recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 486–496, 2024. (Cited on pages 50, 51, 57, and 67.)

[111] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra. Diverse beam search for improved description of complex scenes. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 7371–7379, 2018. (Cited on pages 68 and 75.)

[112] A. Volodkevich, D. Gusak, A. Klenitskiy, A. Pembek, and A. Vasilev. Autoregressive generation strategies for top-k sequential recommendations. *User Modeling and User-Adapted Interaction*, 35(3): 13, 2025. (Cited on page 86.)

[113] M. Wan, D. Wang, J. Liu, P. Bennett, and J. McAuley. Representing and recommending shopping baskets with complementarity, compatibility and loyalty. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1133–1142, 2018. (Cited on page 37.)

[114] J. Wang, H. Lu, Y. Liu, H. Ma, Y. Wang, Y. Gu, S. Zhang, N. Han, S. Bi, L. Baugher, et al. LLMs for user interest exploration in large-scale recommendation systems. In *Proceedings of the 18th ACM International Conference on Recommender Systems*, pages 872–877, 2024. (Cited on page 67.)

[115] L. Wang, J. Zhang, H. Yang, Z.-Y. Chen, J. Tang, Z. Zhang, X. Chen, Y. Lin, H. Sun, R. Song, et al. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37, 2025. (Cited on page 93.)

[116] P. Wang, J. Guo, Y. Lan, J. Xu, S. Wan, and X. Cheng. Learning hierarchical representation model for next basket recommendation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 403–412, 2015. (Cited on pages 30 and 33.)

[117] P. Wang, Y. Zhang, S. Niu, and J. Guo. Modeling temporal dynamics of users' purchase behaviors for next basket prediction. *Journal of Computer Science and Technology*, 34(6):1230–1240, 2019. (Cited on page 33.)

[118] S. Wang, L. Hu, Y. Wang, Q. Z. Sheng, M. Orgun, and L. Cao. Modeling multi-purpose sessions for

next-item recommendations via mixture-channel purpose routing networks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3771–3777, 2019. (Cited on pages 33 and 36.)

[119] S. Wang, L. Hu, Y. Wang, Q. Z. Sheng, M. Orgun, and L. Cao. Intention nets: Psychology-inspired user choice behavior modeling for next-basket prediction. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 6259–6266, 2020. (Cited on page 33.)

[120] W. Wang, H. Bao, X. Lin, J. Zhang, Y. Li, F. Feng, S.-K. Ng, and T.-S. Chua. Learnable item tokenization for generative recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2400–2409, 2024. (Cited on pages 2, 68, 75, and 86.)

[121] Y. Wang, W. Ma, M. Zhang, Y. Liu, and S. Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43, 2023. (Cited on pages 1 and 14.)

[122] P. Williams. Emotions and consumer behavior. *Journal of Consumer Research*, 40(5):viii–xi, 2014. (Cited on pages 32 and 48.)

[123] H. Wu, B. Mitra, C. Ma, F. Diaz, and X. Liu. Joint multisided exposure fairness for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 703–714, 2022. (Cited on pages 30 and 33.)

[124] S. Wu, W. Wei, M. Zhang, Z. Chen, J. Ma, Z. Ren, M. de Rijke, and P. Ren. Generative retrieval as multi-vector dense retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1828–1838, 2024. (Cited on page 92.)

[125] Y. Wu, J. Cao, G. Xu, and Y. Tan. TFROM: A two-sided fairness-aware recommendation model for both customers and providers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1013–1022, 2021. (Cited on pages 30 and 33.)

[126] C. Xu, X. Ye, J. Xu, X. Zhang, W. Shen, and J.-R. Wen. LTP-MMF: Toward long-term provider max-min fairness under recommendation feedback loops. *ACM Transactions on Information Systems*, 43(1):1–29, 2024. (Cited on page 93.)

[127] C. Xu, Z. Deng, C. Rus, X. Ye, Y. Liu, J. Xu, Z. Dou, J.-R. Wen, and M. de Rijke. FairDiverse: A comprehensive toolkit for fairness- and diversity-aware information retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3540–3550, 2025. (Cited on pages 2 and 93.)

[128] C. Xu, C. Rus, Y. Liu, M. de Jonge, J. Xu, and M. de Rijke. Fairness in information retrieval from an economic perspective. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4126–4129, 2025.

[129] C. Xu, C. Rus, Y. Liu, M. de Jonge, J. Xu, and M. de Rijke. Economic perspectives on fairness in information retrieval. In *Proceedings of the 48th European Conference on Information Retrieval*, pages 19–26, 2026.

[130] C. Xu, C. Rus, Y. Liu, M. de Jonge, J. Xu, and M. de Rijke. Towards a responsible web: Economic perspectives on fairness in information retrieval. In *Companion Proceedings of the ACM Web Conference 2026*, pages 23–26, 2026.

[131] K. Yang and J. Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pages 1–6, 2017. (Cited on page 14.)

[132] W. Yang, Y. Pan, C. Gao, S. Zhou, X. Wang, C. Wang, and J. Chen. The pitfall of scaling up: Uncovering and mitigating popularity bias amplification in scaling transformer-based recommenders. In *Proceedings of the 32nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 6044–6055, 2026. (Cited on pages 92 and 93.)

[133] S. Yao and B. Huang. Beyond parity: Fairness objectives for collaborative filtering. *Advances in Neural Information Processing Systems*, 30:2921–2930, 2017. (Cited on page 2.)

[134] X. Ye, C. Xu, Z. Sun, J. Xu, G. Wang, Z. Dong, and J.-R. Wen. LLM-empowered creator simulation for long-term evaluation of recommender systems under information asymmetry. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 201–211, 2025. (Cited on page 93.)

[135] J. Yin, B. Zhu, P. Huo, R. Liu, H. Chen, S. Wang, S. Pan, and C. Zhang. Echoes in filter bubble: Diagnosing and curing popularity bias in generative recommenders. *arXiv preprint arXiv:2605.16825*, 2026. (Cited on page 93.)

[136] Q. Yin, H. Fang, Z. Sun, and Y.-S. Ong. Understanding diversity in session-based recommendation. *ACM Transactions on Information Systems*, 42(1):1–34, 2023. (Cited on pages 30, 31, 33, and 36.)

[137] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan. A dynamic recurrent model for next basket recommendation. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 729–732, 2016. (Cited on pages 2, 13, 16, 18, 30, 33, 40, 49, 56, and 75.)

[138] L. Yu, L. Sun, B. Du, C. Liu, H. Xiong, and W. Lv. Predicting temporal sets with deep neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1083–1091, 2020. (Cited on pages 2, 13, 16, 18, 30, 33, 40, 50, 56, 75, and 85.)

[139] L. Yu, Z. Liu, L. Sun, B. Du, C. Liu, and W. Lv. Continuous-time user preference modelling for temporal sets prediction. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1475–1488, 2023. (Cited on page 2.)

[140] M. Zehlike and C. Castillo. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of the Web Conference 2020*, pages 2849–2855, 2020. (Cited on page 33.)

[141] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. FA*IR: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM International Conference on Information and Knowledge Management*, pages 1569–1578, 2017. (Cited on pages 12 and 14.)

[142] M. Zehlike, K. Yang, and J. Stoyanovich. Fairness in ranking, part II: Learning-to-rank and recommender systems. *ACM Computing Surveys*, 55(6):1–41, 2022. (Cited on page 92.)

[143] J. Zhai, L. Liao, X. Liu, Y. Wang, R. Li, X. Cao, L. Gao, Z. Gong, F. Gu, M. He, et al. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 58484–58509, 2024. (Cited on page 91.)

[144] J. Zhai, Z. Mai, C. Wang, F. Yang, X. Zheng, H. Li, and Y. Tian. Multimodal quantitative language for generative recommendation. In *the 13th International Conference on Learning Representations*, pages 74424–74441, 2025. (Cited on pages 72 and 86.)

[145] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, and X. He. Is ChatGPT fair for recommendation? Evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 993–999, 2023. (Cited on page 93.)

[146] M. Zhang and N. Hurley. Avoiding monotony: Improving the diversity of recommendation lists. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, pages 123–130, 2008. (Cited on pages 30, 33, and 51.)

[147] Y. Zhang, C. Zhang, H. Sun, Q. Qi, L. Zhang, J. Wang, and J. Wang. DiffNBR: Spatio-temporal diffusion with information bottleneck for next-basket recommendation. In *Proceedings of the 19th ACM International Conference on Web Search and Data Mining*, pages 985–995, 2026. (Cited on pages 75 and 86.)

[148] Z. Zhang, J. Zhao, X. Ma, X. Xin, M. de Rijke, and Z. Ren. Cold-starts in generative recommendation: A reproducibility study. *arXiv preprint arXiv:2603.29845*, 2026. (Cited on page 4.)

[149] Y. Zhao, Y. Wang, Y. Liu, X. Cheng, C. Aggarwal, and T. Derr. Fairness and diversity in recommender systems: A survey. *ACM Transactions on Intelligent Systems and Technology*, 16(1):2:1–2:28, 2025. (Cited on pages 30, 34, 46, and 50.)

[150] B. Zheng, Y. Hou, H. Lu, Y. Chen, W. X. Zhao, M. Chen, and J.-R. Wen. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering*, pages 1435–1448, 2024. (Cited on pages 72 and 86.)

[151] Y. Zheng, C. Gao, L. Chen, D. Jin, and Y. Li. DGCN: Diversified recommendation with graph convolutional networks. In *Proceedings of the Web Conference 2021*, pages 401–412, 2021. (Cited on page 36.)

[152] G. Zhou, J. Deng, J. Zhang, K. Cai, L. Ren, Q. Luo, Q. Wang, Q. Hu, R. Huang, S. Wang, et al. OneRec technical report. *arXiv preprint arXiv:2506.13695*, 2025. (Cited on page 92.)

[153] Z. Zhu, J. Wang, and J. Caverlee. Measuring and mitigating item under-recommendation bias in personalized ranking systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 449–458, 2020. (Cited on page 2.)

[154] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, pages 22–32, 2005. (Cited on page 1.)

Summary

Recommender systems are multi-objective systems that connect users with information aligned with their preferences. To evaluate the quality of a recommendation list, we consider not only accuracy metrics but also beyond-accuracy metrics, including item fairness and diversity, as these serve the goals of a sustainable system and support users' long-term engagement. In this thesis, we focus on next basket recommendation (NBR), a special case of sequential recommendation in which the recommendation model handles a sequence of item sets, and users exhibit both repetitive and exploratory purchase behaviors. Over four chapters, we investigate the evaluation protocol and optimization of beyond-accuracy metrics for NBR, spanning from the traditional paradigm to the semantic-ID-based generative paradigm.

First, we measure the item fairness of representative NBR methods using seven fair ranking metrics that cover the principles of both equal opportunity and statistical parity. We then test the sensitivity of these metrics with respect to basket size, position weighting models, and user repeat behavior. Finally, we identify the most robust and applicable metrics for the NBR task.

Second, we propose a two-step repetition–exploration (TREx) framework to verify a short-cut strategy for achieving better accuracy and beyond-accuracy performance in NBR. We further reveal a loophole in the current overall evaluation protocol, which masks the poor accuracy of exploratory recommendation. We therefore refine this protocol by evaluating repetitive and exploratory recommendations separately.

Third, we propose two model-agnostic re-ranking algorithms that jointly mitigate repeat bias and optimize item fairness or diversity. We further extend these algorithms to both unified and combined NBR methods.

Finally, we move to a semantic-ID-based generative paradigm for diverse and exploratory basket recommendation. Specifically, we propose a mask-prompted generative basket recommendation architecture, which uses a bidirectional encoder to encode the basket sequence and a lightweight decoder to autoregressively generate the semantic IDs of the masked item. We further propose a balanced RQ-KMeans tokenizer, which increases codeword assignment entropy to improve downstream recommendation diversity.

Through this series of studies, this thesis advances the evaluation and optimization of beyond-accuracy objectives for NBR, underscoring the importance of distinguishing repetitive from exploratory recommendations during evaluation.

Samenvatting

Aanbevelingssystemen zijn systemen met meerdere doelstellingen die gebruikers koppelen aan informatie die aansluit bij hun voorkeuren. Om de kwaliteit van een aanbevelingslijst te beoordelen, kijken we niet alleen naar nauwkeurigheidsmaatstaven, maar ook naar maatstaven die verder gaan dan nauwkeurigheid, zoals eerlijkheid en diversiteit van items, aangezien deze bijdragen aan de doelstellingen van een duurzaam systeem en de betrokkenheid van gebruikers op de lange termijn ondersteunen. In dit proefschrift richten we ons op *next basket recommendation* (NBR), een speciaal geval van sequentiële aanbevelingen waarbij het model een reeks itemsets verwerkt en gebruikers zowel repetitief als verkennend koopgedrag vertonen. In vier hoofdstukken onderzoeken we het evaluatieprotocol en de optimalisatie van maatstaven die verder gaan dan nauwkeurigheid voor NBR, variërend van het traditionele paradigma tot het semantische, op ID's gebaseerde generatieve paradigma.

Ten eerste meten we de item-eerlijkheid van representatieve NBR-methoden aan de hand van zeven maatstaven voor eerlijke rangschikking die zowel de principes van gelijke kansen als statistische pariteit omvatten. Vervolgens testen we de gevoeligheid van deze maatstaven ten opzichte van de grootte van de *basket*, positiewegingsmodellen en herhalingsgedrag van gebruikers. Ten slotte identificeren we de meest robuuste en toepasbare maatstaven voor de NBR-taak.

Ten tweede stellen we een tweestaps herhaling–verkenning (TREx)-raamwerk voor om een *short-cut*-strategie te verifiëren waarmee een betere nauwkeurigheid en prestaties die verder gaan dan nauwkeurigheid bij NBR kunnen worden bereikt. Verder brengen we een maas in het huidige algemene evaluatieprotocol aan het licht, waardoor de slechte nauwkeurigheid van verkennende aanbevelingen wordt gemaskeerd. Daarom verfijnen we dit protocol door herhalende en verkennende aanbevelingen afzonderlijk te evalueren.

Ten derde stellen we twee modelonafhankelijke algoritmen voor het herschikken van de rangorde voor, die zowel herhalingsvertekening verminderen als de eerlijkheid of diversiteit van de items optimaliseren. We breiden deze algoritmen verder uit naar zowel uniforme als gecombineerde NBR-methoden.

Ten slotte stappen we over op een generatief paradigma op basis van semantische ID's voor diverse en verkennende aanbevelingen van *baskets*. Concreet stellen we een generatieve architectuur voor NBR op basis van maskers voor, die gebruik maakt van een bidirectionele encoder om de volgorde van de *basket* te coderen en een lichtgewicht decoder om de semantische ID's van het gemaskeerde item autoregressief te genereren. Verder stellen we een gebalanceerde RQ-KMeans-tokenizer voor, die de entropie van de codewoordtoewijzing verhoogt om de diversiteit van de aanbevelingen verderop in het proces te verbeteren.

Via deze reeks onderzoeken draagt dit proefschrift bij aan de evaluatie en optimalisatie van doelstellingen die verder gaan dan nauwkeurigheid voor NBR, waarbij het belang wordt benadrukt van het onderscheid tussen repetitieve en verkennende aanbevelingen tijdens de evaluatie.