AI for Personalization: From Predictive to Generative Modeling

Gabriel Bénédict

AI for Personalization: From Predictive to Generative Modeling

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in
de Aula der Universiteit
op vrijdag 01 January 2024, te 14:00 uur

door

Gabriel Bénédict

geboren te Chavannes-Pres-Renens

Promotiecommissie

Promotor:	prof. dr. M. de Rijke	Universiteit van Amsterdam
Co-promotor:	dr. D. Odijk	RTL NL
Overige leden:	prof. dr. L. Name	Universiteit van Amsterdam
	prof. dr. L. Name	Universiteit van Amsterdam
	prof. dr. L. Name	Universiteit van Amsterdam
	prof. dr. L. Name	Universiteit van Amsterdam
	prof. dr. L. Name	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

This research was (partially) funded by Bertelsmann SE & Co. KGaA.

Cover by Off Page, Amsterdam
Printed by Off Page, Amsterdam

ISBN:

Certains pensent qu'ils font un voyage, en fait, c'est le voyage qui vous fait ou vous défait.

– Nicolas Bouvier

Acknowlegements

Gabriel Bénédict
Amsterdam
January 2024

Table of Contents

Introduction

Video streaming platforms have changed the way people consume and interact with digital media [10]. One of the key innovations of video streaming platforms with regards to traditional television, is their ability to tailor the experience to each single user, given their past behavior on the platform (a.k.a personalization [5, 27]). In other words, the *user journey* (the user's perspective) is steered by the platform's algorithms.

The user journey consists of several steps. First, users come to the platform with some *intents* (e.g., binge-watching a series, finding a family-friendly movie, discovering new genres, etc.) [3]. Then, they see a customized home page with various horizontal *recommendation* strips (see Figure 1.1). Each strip contains videos with (sometimes personalized) *thumbnails* (the clickable image that represents the video content). Over time, users interact with the platform and leave *behavioral* signals (e.g., clicks, watch time, bookmarks, ratings, etc.). From the platform's perspective, deciphering how these behavioral signals, prompted by user intents, translate into overall *satisfaction* remains a complex challenge.

[Gab: not sure if this paragraph on the history of personalization is needed]

Behavioral signals are used to target users individually: personalize the experience. The term personalization has been used for long to describe this strategy [5, 27]. But historically personalization has been used as an umbrella term for approaches more akin to targeting user segments / groups than actually serving each single user differently (as the term personalization seems to hint at). At first, in the domain of e-commerce in the late 90s, personalization was restricted to email newsletters and aimed at user groups [15]. Personalization then appeared on early video streaming platforms like Netflix in the form of one dimensional lists (a.k.a. recommendation strip) [15]. Now personalization is in the ordering of the strip (multiple 1-dimensional lists), the thumbnail, in the font title of the thumbnail, in search [10], etc.

The personalization pipeline is the accumulation of a platform's algorithms for a tailored user journey. One part of the pipeline retrieves data that feeds all other algorithms: collecting user data and analyzing user behavior [23]. The data granularity can range from the number of items watched (just one data point per user session), all the way to recordings of all mouse movements (thousands of data points per session). With that session-level data, streaming platforms

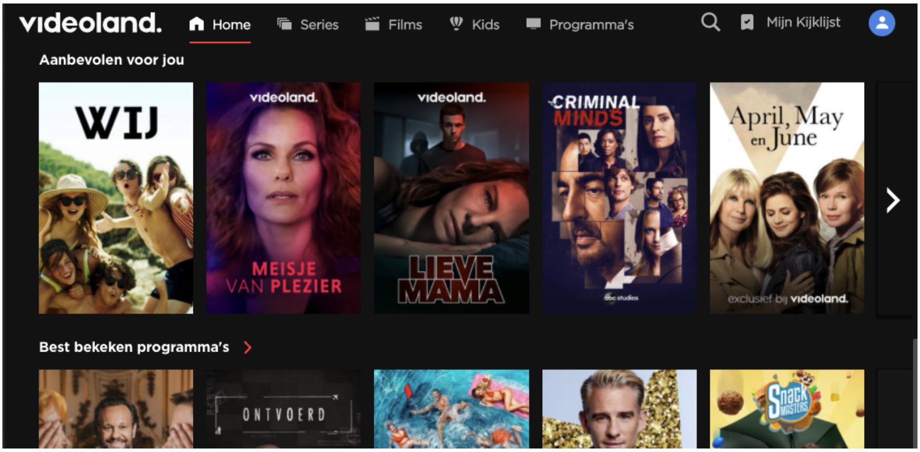

Figure 1.1: Videoland's Recommendation strips

attempt to predict what the user will do and adapt the user journey to the user: the next movie to watch, the subsequent logins, the midterm satisfaction (typically, the amount of time spent on the platform per month), all the way to churn rate (subscription cancellation rate) [12]; from short to long term forecast. These predictive models are tested first on a simulated platform with simulated users 'offline'. Models are then evaluated on the platform exposed to users 'online' over several iterations and over time. In the evaluation phase, preferences, and interaction patterns are captured again as a feedback loop [1, 10]. Aside from the measurable user feedback signals (such as clicks, watch time, time on the platform etc.), a platform can also take hidden signals into consideration. In this thesis, we give some attention to user intent (watching a next episode on a favorite show, looking for new content, bookmarking items for later viewing, etc.). Besides satisfying the users, the platform also bears a responsibility, because it is able to steer the user towards certain consumption behaviors. For example, the platform has to ensure that the content it offers is *diverse* and promotes representative voices (e.g., promoting screenwriters of different genres, movies in different languages, a variety of movie genres).

For the purpose of this thesis, we will call this combination of algorithms, the *personalization pipeline* (the platform's perspective). This pipeline is geared towards simple metrics like the number of minutes seen [20] and the churn rate [17], but is linked with a responsibility to balance longer term user satisfaction [12], content diversity, and other ethical considerations [11].

In this thesis, we propose individual tools that map to the user journey above, for the steps of *recommendations*, *thumbnail* selection, *intent-satisfaction* linking and *diversity* measurement. Together, these tools form our proposed **personalization pipeline**. By adapting diffusion models [25] from the image domain (continuous-2D-structured-data) to recommendation (binary-1D-unstructured-data), we open the door to the use of priors in recommendation (preferred movie genres, past behavior or incomplete viewing history, etc.), like is seen in diffusion for images (an image description, a previous reference image, a masked image, etc.). Diffusion models are physics inspired neural models, that include a forward (perturbation) and backward (learning) process on each example [25]. As for personalized *thumbnails*, existing research in multilabel image classification is highly reliant on variations of the binary cross-entropy loss [8], but we think that the multilabel setting (as opposed to the binary or multiclass setting) requires its own solution. For the next step, we identified a lack of a systematic approach for *intent-satisfaction* studies, that would provide survey design, code and modern bayesian approaches to the problem. Finally, we could not find a *diversity* metric for news / movies recommendations that is distribution agnostic – to adapt to any distributions of discrete normative standpoints – and rank-aware – to accommodate for the propensity of a user to scroll up to an item on a ranked list.

In short, we focus on the video streaming platform ecosystem, exploring the challenges and opportunities of personalization, recommendation, and user behavior analysis. By combining survey methods, modeling, adaptive testing, and behavioral analysis, this study aims to contribute to the development of video streaming platforms that can provide user satisfaction in a responsible way.

1.1 Research Outline and Questions

We scope the thesis around four research questions, each corresponding to a research chapter in the thesis.

Personalization on streaming platforms is often seen as a way of predicting what users want to watch based on their preferences and behavior. Personalization can also be seen as a creative process (see recent generative recommendation models [13, 19, 21]) with ethical impact (see recent literature on responsible recommendation [11, 31, 33]) that involves generating new content and experiences for users through a pipeline.
[MdR: What are the assumptions that you are making? What is the context?] **[Gab: what kind of assumptions are expected here? I can think of a lot of assumptions regarding the model/data/platform/pipeline?]**

Our first research question addresses the entry point of the user journey on a personalized platform, namely recommendations on the home page (the page where a user lands first on a streaming platform).

RQ1 Can we use diffusion to do recommendation in the classical user-item matrix setting?

Traditionally, recommender systems directly retrieve content from the library to the user. Alternatively, user instructions and feedback are fed to a generator of personalized content, before retrieving and ranking from that new library of generated content, according to the recent generative recommendation paradigm [32]. According to its instigators, the paradigm covers individual content generated from scratch (like diffusion based image creation) or a recommendation of content that is created in a generative way (Large Language Models (LLMs) for recommendation). Somewhat combining the two concepts, we investigate how diffusion models can be used to generate a list of recommended content. Diffusion has been applied to images, music and other modalities. Unlike these, the classical recommendation setting of the user-item matrix [16] does not entail spatial relationships between data points: contrary to pixels on an image, there is no information encoded in the allocation of users and item on a matrix. We illustrate this in RecFusion [7], where we first use Unets [24] to fit data in a spatial way, before going back to the classical recommendation neural setting of feeding data user-by-user. For this one-dimensional user vector, we provide a proof and first experiments to show that a binomial (Bernoulli) diffusion process is viable.

After recommendation, the next step of our pipeline caters to the display of recommended videos via thumbnails.

RQ2 Is there a way we can generate personalized thumbnails for each item on a streaming platform?

Personalization can be seen at different levels of granularity: from targeting user segments (into interests, age groups, etc.) to targeting single users differently. For this research question, we are interested in how thumbnails (i.e. static images) can be classified into different categories, more than knowing if we can target each single user. We therefore opt for a least granular option: we assume that each user has a favorite genre. We can provide a thumbnail personalized to that genre (e.g., show a romantic scene from an action movie, if the favorite genre is romance). Given editorial or automatically selected candidates for thumbnails, we wish to display the one that is most closely associated with the user's preferences. This reduces to a multilabel classification problem: given an image, predict one, or many, genre(s). When thinking about classification, the confusion matrix [26] – with its false positive, true positive, false negative and true positive quadrants – is a classic way to build evaluation metrics. But these metrics are hardly used at training time. We think it is because these quadrant values require counting, which is not differentiable at training time for gradient descent [14, 22]. We propose a way to build surrogates to these count metrics via sigmoid functions. More precisely, we look at maximizing for the F1 score via our sigmoidF1 surrogate loss function [2], as a multilabel classification loss over an entire batch. We show that this improves on classical image and text benchmarks with classical backbones (CNN [9] and transformers [29]).

Recommendations and thumbnails are what primes the users interactions with the platform. This relates to the next step in our pipeline.

RQ3 Are users' intents together with their behavioral data useful signals to predict or explain satisfaction on a video streaming platform?

Streaming platforms have access to user implicit (clicks, scrolls, time on the platform, etc.) and explicit (ratings and bookmarks) feedback via their personalization pipeline. Some of the user behaviors will remain forever hidden from the platform though for privacy or technical reasons (e.g., how many people sit in front of the device, the content the user consumes on other platforms). Among them, we explore user intents of a video streaming platform. Previous work has defined intents for music [20]; we propose to define them for video and propose a transparent approach by revealing our survey design, code and simulated data. In [20], logistic regression was used to predict satisfaction based on intents and behavioral data. We propose to use random forests and Bayesian hierarchical modeling to enhance accuracy and interpretability respectively.

Finally, we close off our pipeline with a responsible approach to diversity.

RQ4 Can we formulate a divergence metric that measures the normative diversity of recommendations?

Videos and especially news platforms serve content that is opinionated. Over time, platforms have been growing their engineering teams to cover more and more of the user journey stages (home page, title fonts, watch/read next etc.) [10], with more and more powerful and sometimes generative models. The user is thus influenced by the platform's algorithm and thus the platform's explicit or implicit norms and values. Can we empower a video/news platform to measure its ability to cater to its norms and values? We would like to account for how a platform means to properly inform citizens (as defined by [11]) and any form of diversity metric (topic, presence of alternative voices, complexity of the text, etc.). RADio [30], the framework we propose caters to these normative aspects but also to the specific recommendation context: RADio is rank aware and caters for any kind of discrete distribution via a our proposed rank-aware Jensen-Shannon divergence [18]. This chapter is focused on news recommendation but trivially generalizes to any domain that has categories (e.g., video streaming with movie genres).

Our research questions have been outlined in this section. The main contributions of this thesis will be summarized in the next section.

1.2 Main Contributions

In this section, we summarize the main contributions of this thesis. We separate theoretical from artifact (that is, tool and experimental design) contributions.

Theoretical Contributions

- An adaptation of diffusion to unstructured data, where there is no spatial dependency (Chapter 2).
- The use of diffusion for binary and/or 1D data: A demonstration that KL divergence is also suited for binary data and that the Bernouilli Markov process has the same properties as its Gaussian counterpart (Chapter 2).
- A multilabel loss function that accounts for all examples in a batch (Chapter 3).
- An F1 score surrogate as a loss function (Chapter 3).

- An account of the current limitations and underreporting of thresholding at inference time (Chapter 3).
- A proposal of typical intents for a video streaming that we divide into explorative and decisive categories (Chapter 4).
- A diversity metric that adapts to any normative concept and expressed as the divergence between two (discrete) distributions, rank-aware and mathematically grounded in distributional divergence statistics (Chapter 5).

Artifact Contributions

- A frequentist logistic regression model, we test Bayesian multilevel models for visualization and explanations, along with random forests for improved accuracy (Chapter 4).
- A reproducibility study from music to video streaming platforms of intent-satisfaction modeling (this time with code and synthetic data) (Chapter 4).
- An in-app survey design for a medium size streaming platform ($\sim$1 million users) and corresponding synthetic data (Chapter 4).
- A metadata enrichment pipeline (e.g., sentiment analysis, named entity recognition) to extract normative concepts from news articles (Chapter 5).

1.3 Thesis Overview

This first chapter introduces the main topics and goals of this thesis, and suggests some possible ways to read it. The thesis has six chapters in total, and this is the first one. The following four chapters each address one of the research questions that we presented in Section 1.1. Each chapter is based on a paper (see Section 1.4 below), and can be read on its own. If the reader is interested in the entire thesis, we recommend following the original order of chapters, as they follow the *user journey* and its respective *personalization pipeline* on a streaming platform. The final chapter summarizes the main findings and contributions of this thesis, and proposes some future research directions.

1.4 Origins

We list the publications that are the origins of each chapter below.

Chapter 2 is based on the following paper:

- Gabriel Bénédict, Olivier Jeunen, Samuele Papa, Samarth Bhargav, Daan Odijk, and Maarten de Rijke. RecFusion: A Binomial Diffusion Process for 1D Data for Recommendation, *arXiv.* 2306.08947, 2023.

GB wrote the first draft, code, mathematical formulations, designed and ran experiments. GB was helped all along the way via discussion with the coauthors. Coauthors then edited the first draft together with GB. GB did most of the writing.

Chapter 3 is based on the following paper:

- Gabriel Bénédict, Hendrik Vincent Koops, Daan Odijk, and Maarten de Rijke. sigmoidF1: A Smooth F1 Score Surrogate Loss for Multilabel Classification, *Transactions of Machine Learning Research*. 2022.

GB wrote the first draft, code, mathematical formulations, designed and ran experiments. GB was helped all along the way via discussion with the coauthors. Coauthors then edited the first draft together with GB. GB did most of the writing.

Chapter 4 is based on the following paper:

- Gabriel Bénédict, Daan Odijk, and Maarten de Rijke. Intent-Satisfaction Modeling: From Music to Video Streaming, 1(3), Art. 13, 2023.

GB wrote the first draft, code, mathematical formulations, designed and ran experiments. GB was helped all along the way via discussion with the coauthors. Coauthors then edited the first draft together with GB. GB did most of the writing.

Chapter 5 is based on the following paper:

- Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, and Maarten de Rijke. RADio – Rank-Aware Divergence Metrics to Measure Normative Diversity in News Recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems*, page 208–219. 2022.

GB, together with SV, wrote the first draft, code, mathematical formulations, designed and ran experiments. GB was helped all along the way via discussion with the coauthors. Coauthors then edited the first draft together with GB. GB and SV did most of the writing.

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

- Garbiel Bénédict, Ruqing Zhang, and Donald Metzler. Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 3460–3463. 2023.
- Ali Vardasbi, Gabriel Bénédict, Shashank Gupta, Maria Heuss, Pooya Khandel, Ming Li, and Fatemeh Sarvi. The University of Amsterdam at the TREC 2021 Fair Ranking Track, *TREC Fair Ranking*. 2021.
- Gabriel Bénédict. Generative Adversarial Networks, *Spectra ML Review Paper Competition*. 09.00009, 2021.

Generative Recommendations with Diffusion

RQ1: Can we use diffusion to do recommendation in the classical user-item matrix setting?

Reproducibility

To facilitate the reproducibility of the work in this chapter, our code is available at `https://github.com/gabriben/recfusion`.

This chapter is under submission at International Conference on Learning Representations (ICLR) under the title "RecFusion: A Binomial Diffusion Process for 1D Data for Recommendation" [?].

Chapter 3

Metrics as Losses

In this chapter, we address the following research question:

RQ2: Is there a way we can generate personalized thumbnails for each item on a streaming platform?

Reproducibility

To facilitate the reproducibility of this work, our code is available at `https://github.com/gabriben/metrics-as-losses`.

This chapter was published at the Transactions of Machine Learning Research (TMLR) under the title "sigmoidF1: A Smooth F1 Score Surrogate Loss for Multilabel Classification" [2].

Chapter 4

Intent, Behavior and Satisfaction

In this chapter, we address the following research question:

RQ3: Are users' intents together with their behavioral data useful signals to predict or explain satisfaction on a video streaming platform?

Reproducibility

To facilitate the reproducibility of the work in this chapter, our code is available at `https://github.com/rtlnl/streaming-intent-model`.

This chapter was published at the ACM Transactions on Recommender Systems (TORS) under the title "Intent-Satisfaction Modeling: From Music to Video Streaming" [3]

Chapter 5

Normative Diversity

In this chapter, we address the following research question:

RQ4: Can we formulate a divergence metric that measures the normative diversity of recommendations?

Reproducibility

To facilitate the reproducibility of the work in this chapter, our code is available at `https://github.com/svrijenhoek/RADio`.

This chapter was published at the ACM Conference on Recommender Systems (RecSys 2022) under the title "RADio – Rank-Aware Divergence Metrics to Measure Normative Diversity in News Recommendations" [30], , where it won a best paper runner up award.

Chapter 6

Conclusions

In this thesis...

6.1 Main Findings

In this section, we describe our main findings across the three parts of the thesis.

The first part of this thesis focused on a generative personalization pipeline throughout the user journey on a video streaming platform. Along this pipeline, we first present RecFusion, a system that uses diffusion models to generate novel and relevant recommendations for users, as part of the emerging field of Generative Information Retrieval. To make these recommendations more appealing, we also propose a method to generate personalized stills from movies using sigmoidF1, a technique that adapts the image quality and style to the user's taste. We analyze how the user interactions on streaming platforms are influenced by not only the explicit data that is collected by web analytics, but also the implicit data that is hidden from them, using our intent-satisfaction framework. Finally, we ensure that the recommendations we generate respect the normative diversity of the users and the content providers, using RADio, a framework that measures and optimizes the fairness and diversity of the recommendations.

In Chapter 2, we asked our first research question:

RQ1 Can we use diffusion to do recommendation in the classical user-item matrix setting?

The answer to **RQ1** is yes:

In Chapter 3, we then turned to our next research question:

RQ2 Is there a way we can generate personalized thumbnails for each item on a streaming platform?

The answer to **RQ2** is yes:

In Chapter 4, we investigated the following research question:

RQ3 Are users' intents together with their behavioral data useful signals to predict or explain satisfaction on a video streaming platform?

The answer to **RQ3** is yes:

We asked our final research question in Chapter 5:

RQ4 Can we formulate a divergence metric that measures the normative diversity of recommendations?

We answered **RQ4** by ...

6.2 Future Directions

In this section, we describe some limitations of the methods proposed in this thesis and identify potential avenues for future work.

Final thoughts

Bibliography

[1] Joeran Beel and Stefan Langer. A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research-Paper Recommender Systems, page 153–168, *Lecture Notes in Computer Science*. 2015. (Cited on page 2.)

[2] Gabriel Bénédict, Hendrik Vincent Koops, Daan Odijk, and Maarten de Rijke. sigmoidF1: A Smooth F1 Score Surrogate Loss for Multilabel Classification, *Transactions of Machine Learning Research*. 2022. (Cited on pages 4 and 11.)

[3] Gabriel Bénédict, Daan Odijk, and Maarten de Rijke. Intent-Satisfaction Modeling: From Music to Video Streaming, 1(3), Art. 13, 2023. (Cited on pages 1 and 13.)

[4] Garbiel Bénédict, Ruqing Zhang, and Donald Metzler. Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 3460–3463. 2023.

[5] Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. Modeling the Impact of Short- and Long-Term Behavior on Search Personalization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 185–194. 2012. (Cited on page 1.)

[6] Gabriel Bénédict. Generative Adversarial Networks, *Spectra ML Review Paper Competition*. 09.00009, 2021.

[7] Gabriel Bénédict, Olivier Jeunen, Samuele Papa, Samarth Bhargav, Daan Odijk, and Maarten de Rijke. RecFusion: A Binomial Diffusion Process for 1D Data for Recommendation, *arXiv*. 2306.08947, 2023. (Cited on page 4.)

[8] Ronald A. Fisher. On an Absolute Criterion for Fitting Frequency Curves, 41:155–160, *Messenger of Mathematics*. 1912. (Cited on page 3.)

[9] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, 36:193 – 202, *Biological Cybernetics*. 1980. (Cited on page 4.)

[10] Carlos A. Gomez-Uribe and Neil Hunt. The Netflix Recommender System, 6(4):1–19, 2015. (Cited on pages 1, 2, and 5.)

[11] Natali Helberger. On the Democratic Role of News Recommenders, 7(8):993–1012, *Digital Journalism*. https://doi.org/10.1080/21670811.2019.1623700, 2019. (Cited on pages 2, 3, and 5.)

[12] Henning Hohnhold, Deirdre O'Brien, and Diane Tang. Focusing on the Long-term (It's Good for Users and Business), 2015. (Cited on page 2.)

[13] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large Language Models are Zero-Shot Rankers for Recommender Systems. 2305.08845, 2023. (Cited on page 3.)

[14] J. Kiefer and J. Wolfowitz. Stochastic Estimation of the Maximum of a Regression Function, 23(3):462 – 466, *The Annals of Mathematical Statistics*. 1952. (Cited on page 4.)

[15] Ron Kohavi and Foster Provost. Applications of Data Mining to Electronic Commerce, page 5–10, *Applications of Data Mining to Electronic Commerce*. 2001. (Cited on page 1.)

[16] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix Factorization Techniques for Recommender Systems, 42(8):30–37, 2009. (Cited on page 4.)

[17] Jae Sik Lee and Jin Chun Lee. Customer Churn Prediction by Hybrid Model. In Xue Li, Osmar R. Zaïane, and Zhanhuai Li, editors, *Advanced Data Mining and Applications*,

pages 959–966. 2006. (Cited on page 2.)

[18] J. Lin. Divergence measures based on the Shannon entropy, 37(1):145–151, *IEEE Transactions on Information Theory*. 1991. (Cited on page 5.)

[19] Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. ONCE: Boosting Content-based Recommendation with Both Open- and Closed-source Large Language Models. 2305.06566, 2023. (Cited on page 3.)

[20] Rishabh Mehrotra, Mounia Lalmas, Doug Kenney, Thomas Lim-Meng, and Golli Hashemian. Jointly Leveraging Intent and Interaction Signals to Predict User Satisfaction with Slate Recommendations. In *The World Wide Web Conference*, page 1256–1267. 2019. (Cited on pages 2 and 5.)

[21] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan H. Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. Recommender Systems with Generative Retrieval. 2305.05065, 2023. (Cited on page 3.)

[22] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method, 22(3):400 – 407, *The Annals of Mathematical Statistics*. 1951. (Cited on page 4.)

[23] Royi Ronen, Elad Yom-Tov, and Gal Lavee. Recommendations meet web browsing: enhancing collaborative filtering using internet browsing logs, 2016. (Cited on page 1.)

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. page 234–241. Springer International Publishing, 2015. (Cited on page 4.)

[25] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, 2015. (Cited on pages 2 and 3.)

[26] Stephen V. Stehman. Selecting and interpreting measures of thematic classification accuracy, 62(1):77–89, *Remote Sensing of Environment*. 1997. (Cited on page 4.)

[27] Jaime Teevan, Susan T. Dumais, and Eric Horvitz. Personalizing Search via Automated Analysis of Interests and Activities. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 449–456. 2005. (Cited on page 1.)

[28] Ali Vardasbi, Gabriel Bénédict, Shashank Gupta, Maria Heuss, Pooya Khandel, Ming Li, and Fatemeh Sarvi. The University of Amsterdam at the TREC 2021 Fair Ranking Track, *TREC Fair Ranking*. 2021.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. (Cited on page 4.)

[30] Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, and Maarten de Rijke. RADio – Rank-Aware Divergence Metrics to Measure Normative Diversity in News Recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems*, page 208–219. 2022. (Cited on pages 5 and 15.)

[31] Sanne Vrijenhoek, Lien Michiels, Johannes Kruse, Jordi Viader Guerrero, Alain Starke, and Nava Tintarev. NORMalize: The First Workshop on Normative Design and Evaluation of Recommender Systems. In *17th ACM Conference on Recommender Systems*. 2023. (Cited on page 3.)

[32] Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. Generative Recommendation: Towards Next-generation Recommender Paradigm. 2304.03516v1, 2023. (Cited on page 3.)

[33] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation. In *17th ACM Conference on Recommender Systems*. 2023. (Cited on page 3.)

Summary

Samenvatting