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

Chapter 1

Introduction

check these https://dl.acm.org/doi/abs/10.1145/3460231.3474612#sec-ref [citation needed]

Video streaming platforms have changed the way people consume and interact with digital media [citation needed]. Using machine learning to deliver personalized and recommended content to users based on their behavior and preferences is not something new [3, 7]. At first, personalization was restricted to email newsletters, then appeared on the platform in a form of 1 dimensional lists [citation needed]. Now personalization is in the ordering of the strip, the thumbnail, in the font title of the thumbnail, in search [citation needed]. The outcome is an entire *user journey* (the user's perspective) steered by the platforms algorithms. 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 KPIs like number of minutes seen [citation needed] and churn rate [citation needed], but is linked with a responsibility to balance longer term user satisfaction, content diversity, and ethical considerations [citation needed].

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.) [2]. Then, they see a customized home page with various horizontal *recommendation* strips (see Figure 1.1). Each strip contains videos with (sometimes personalized) *thumbnails*. 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 behaviors, prompted by user intents, translate into overall *satisfaction* remains a complex challenge.

The personalization pipeline is our description of the accumulation of a platform's algorithms that cater for a better user journey. The pipeline includes collecting user data, analyzing user behavior. These signals, also called user feedback are often provided by an external analytics business and the signal granularity is limited by the amount of data and how much a streaming platform is willing to pay [citation needed]. It can range from number of item 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 attempt to predict what the user will do, to adapt the user journey to the user: the next movie to watch, the subsequent logins, the

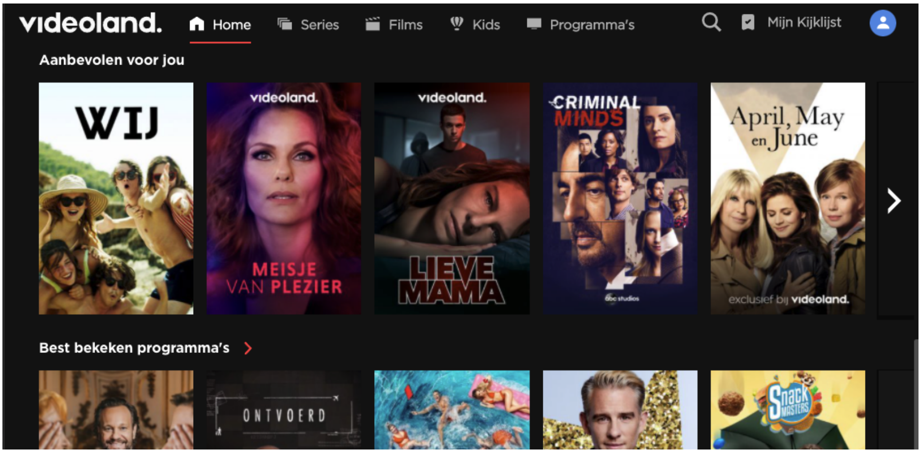

Figure 1.1: Videoland's Recommendation strips

midterm satisfaction, all the way to churn; in increasing order of prediction horizon. These predictive models are tested first offline and then evaluated online over several iterations and over time. In the evaluation phase, preferences, and interaction patterns are captured again as a feedback loop. Aside from the measurable user feedback signals, 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 viewership). Besides satisfying the users, the platform also has to ensure that the content it offers is *diverse* and promotes representative. For example, promoting screenwriters of different genres, movies in different languages, a variety of movie genres. These concerns are revisited below.

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**. For *recommendations*, we found a good amount of literature on diffusion models [] for continuous 2D structured data but little-to-no literature on any relaxation of the above. We propose to explore formulation of diffusion models for 1D binary unstructured data. As for personalized *thumbnails*, existing research in multilabel image classification is highly reliant on variations of the binary cross-entropy loss [], 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 that is distribution agnostic and rank-aware, to adapt to any normative standpoints of a platform issuing news / movies recommendations.

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, content 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 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 and ethical process that involves generating new content and experiences for users through a pipeline. Our first research question addresses the entry point of the user journey on a personalized platform, namely recommendations on the home page.

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

Assuming recommendations are to be linked with a certain form of creativity, we harness the power of generative models. The emergent diffusion models field has been applied to images, music and other modalities. Unlike these, the classical recommendation setting of the user-item matrix 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, where we first use Unets to fit data ina spatial way, before going back to the classical recommendation neural setting of feeding data user-by-user. For this one-dimensional user vector, ee propose 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 these recommended videos via thumbnails.

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

Taking the simplifying assumption 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 this most closely associated with the user's preferences. This reduces to a multilabel classification problem: given an image, predict one, or many, genre(s) they associate with. With sigmoidF1 we propose a surrogate F1 Score as a loss function, as a multilabel classification loss. At training time, we approximate step functions (i.e. confusion matrix counts: true positives, false positives etc.) with sigmoid functions and calculate an F1 score over each entire batch. We show that this improves on classical image and text benchmarks with classical backbones (CNN and transformers). 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?

By merging user behavior on the website and user surveys, we connect implicit and explicit user feedback and link them to satisfaction. We reproduce a study [**?**], but this time propose a transparent approach by revealing our survey design, code and simulated data. We use Bayesian multilevel modeling to reveal the relationships between intents, behavioral data and satisfaction. 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?

Can we empower a video/news platform to measure its ability to cater to its norms and values? We would like to account for any form of democratic norms (how a platform means to properly inform citizens) and any form of diversity metric (topic, presence of alternative voices, complexity of the text, etc.). The framework we propose caters to these normative aspects but also to the specific recommendation context: RADio is rank aware and caters any kind of discrete distribution via a our proposed rank-aware Jensen Shannon divergence. This work 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 (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 (∼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*, 2023, 2306.08947.

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. *ACM Transactions On Recommender Systems*, 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:

- Gabriel Bénédict, Ruqing Zhang, and Donald Metzler. The First Workshop on Generative Information Retrieval. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 2023, 2306.02887.

- 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*, 2021.

Chapter 2

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" [1].

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" [2]

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" [9], , 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] 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 page 9.)

[2] Gabriel Bénédict, Daan Odijk, and Maarten de Rijke. Intent-Satisfaction Modeling: From Music to Video Streaming, 1(3), Art. 13. *ACM Transactions On Recommender Systems*, 2023. (Cited on pages 1 and 11.)

[3] 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.)

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

[5] 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*, 2023, 2306.08947.

[6] Gabriel Bénédict, Ruqing Zhang, and Donald Metzler. The First Workshop on Generative Information Retrieval. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 2023, 2306.02887.

[7] 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.)

[8] 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.

[9] 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 page 13.)

Summary

Samenvatting