氏　　名　田沼 巌

学位(専攻分野)　博士(統計科学)

学 位 記 番 号　総研大甲第 2450 号

学位授与の日付　2023 年 9 月 28 日

学位授与の要件　複合科学研究科　統計科学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Discrete Distribution-based Collaborative Filtering Utilizing Various Data Sources for Recommendation Systems

論 文 審 査 委 員　主　査　持橋 大地
統計科学コース 准教授
日野 英逸
統計科学コース 教授
Le Thanh Tam
統計科学コース 助教
松井 知子
統計科学コース 教授
後藤 真孝
産業技術総合研究所
人間情報インタラクション研究部門 首席研究員

# 博士論文の要旨

氏　　名：田沼　巌

論文題目：
Discrete Distribution-based Collaborative Filtering Utilizing Various Data Sources for Recommendation Systems

The growth of web services and applications has enabled access to a huge variety of content, such as articles, music, movies, and games. Suppose a huge amount of information is now accessible. In that case, this means that the desired information must be found from the vast amount of data, i.e., searching for the desired information has become more complex. Recommendation systems, a system that selects objects, information, or products considered valuable to the user and present them, have become an essential technology for various services, not limited to e-commerce sites and streaming services, but are used in various services.

This research investigates extensions to collaborative filtering, a typical approach for recommendation systems. The common challenge in recommendation systems is that relevance data between users and items, which is both observational data and objective variables, is very sparse, and to overcome this challenge, utilizations of various external data are investigated. In this research, we investigate the utilization of external data, particularly for collaborative filtering assuming discrete distributions. A number of collaborative filtering extensions have been considered, often Gaussian distribution-based. In contrast, the data treated in recommendation systems are often discrete variables, such as binary variables indicating whether users are interested or not, ratings of items, and the counts of usage. Handling such data with discrete distributions has advantages in terms of interpretability, and it would be beneficial to investigate extensions specific to the case of assuming a discrete distribution. Concretely, we investigate the utilization of content information with Poisson distribution-based collaborative filtering and the utilization of proportion information of ratings with binomial distribution-based collaborative filtering.

The former is the investigation of the utilization of content information, i.e., hybrid recommendation, with Poisson distribution-based collaborative filtering. Hybrid recommendation, based on collaborative filtering and supplemented with auxiliary content information, is being actively researched due to its ability to overcome the cold-start problem. Many proposed hybrid methods make recommendations using Gaussian distribution-based collaborative filtering even though they handle variables that tend

to be non-Gaussian, such as the number of interactions. We present a technique that uses a hybrid recommendation framework based on collaborative filtering that models the number of interactions as a Poisson-distributed and variational autoencoder-based content information generation process that shares latent variables with collaborative filtering. As a prior for the shared latent variables, we use a gamma distribution, which is a conjugate prior of a Poisson distribution. An implicit-derivative-based reparameterization trick enables the use of a gamma distribution in a variational autoencoder. The latent variables in the generative model are inferred using the stochastic gradient variational Bayes approach, taking the number of interactions corresponding to users and items and content information as input. In accordance with the inference, unobserved interactions between users and items are predicted for recommendation. Using a neural network-based generative model for content information enables the framework to handle various types of content information. Experimental results show that the proposed method utilizes content information effectively for predicting the number of interactions and that it should aid in overcoming the cold-start problem.

The latter is the investigation of bias reduction utilizing the proportion of ratings for collaborative filtering based on binomial distribution under exposure bias. Treatment of biases in observed data is one of the major challenges in many statistical and machine learning applications. It is not an exception of them in the context of recommendation systems, and various methods based on causal inference are being investigated. We investigate a collaborative filtering technique that robustly predicts ratings from biased observation. Utilizing the unbiased proportion of rating in the different data sources, we extend collaborative filtering by adding a term for the proportion of rating estimates from collaborative filtering to be closer to the proportion of unbiased data. The proposed method is based on collaborative filtering with binomial matrix factorization. By treating the proportion of ratings in the unbiased case as a probability distribution and adding a constraint term that minimizes the KL divergence with the estimates by collaborative filtering, the proposed method aims to obtain less biased estimates from observations that include bias. The binomial distribution-based collaborative filtering allows direct comparison between unbiased proportions and the proportions estimated from collaborative filtering since the observations can be treated as discrete variables. We show that the proposed method is also applicable to extensions for collaborative filtering techniques with causal inference approaches such as inverse propensity weighting matrix factorization. Experimental results show that the implementation of the proposed method alone and its combination with inverse propensity score weighting matrix factorization improved the performance of recommendations.

（別紙１ Separate Sheet 1）

**Results of the doctoral thesis defense**

# 博士論文審査結果

Name in Full
氏 名　　田沼　巌

Title
論文題目　　Discrete Distribution-based Collaborative Filtering Utilizing Various Data Sources for Recommendation Systems

出願者は、2023 年 8 月 21 日 17 時半から 1 時間にわたって、審査委員 5 名、聴講者 1 名、出願者が参加のもとで博士論文公聴会を行った。博士論文は受理された論文[1]と論文[2]に基づいている。


[1] I. Tanuma and T. Matsui, "Variational Autoencoder-Based Hybrid Recommendation With Poisson Factorization for Modeling Implicit Feedback." IEEE Access, DOI:10.1109/ACCESS.2022.3180051, 2022.

[2] I. Tanuma and T. Matsui, "Rating Proportion-aware Binomial Matrix Factorization for Collaborative Filtering," IEEE Access, DOI: 10.1109/ACCESS.2023.3303322, 2023.


博士論文の内容は次の通りである。

近年、Web サービスやアプリケーションの発展により、記事、音楽、映画、ゲームなど、膨大な種類のコンテンツにアクセスできるようになった。そのため膨大な量の情報にアクセスできるようになった場合、膨大なデータから目的の情報を探し出さなければならず、目的の情報を探すことがより複雑になってきた。レコメンデーションシステムは、ユーザーにとって価値があると思われる対象や情報、商品を選んで提示するシステムで、e-コマースサイトやストリーミングサービスに限らず、様々なサービスで利用されている必須技術となっている。

本研究では、推薦システムの代表的なアプローチである協調フィルタリングの拡張を検討している。推薦システムにおける共通の課題は、観測データであり目的変数でもあるユーザーとアイテムの関連性データが非常に疎であることである。この課題を克服するために、これまで様々な外部データの活用が検討されてきた。本研究では、特に離散分布を仮定した協調フィルタリングにおける外部データの活用を検討している。これまで協調フィルタリングの拡張は数多く検討されてきたが、多くはガウス分布に基づくものであった。一方、推薦システムで扱われるデータは、ユーザの興味の有無を示す二値変数、アイテムの評価、利用回数など、離散的な変数であることが多い。このようなデータを離散分布で扱うことは解釈のしやすさの点で有利であり、離散分布を仮定した場合に特化した拡張を検討することは有益であると考えられる。そのため本研究では、（１）ポアソン分布に基づく協調フィルタリングによるコンテンツ情報の活用、（２） 二項分布に基づく協調フィルタリングによる評価の割合情報の活用について検討している。

（１）については、ポアソン分布に基づく協調フィルタリングによってコンテンツ情報を活用するハイブリッド型の推薦システムについて検討している。従来、協調フィルタリ

ングをベースに、補助的なコンテンツ情報で補完するハイブリッド推薦が、コールドスタート問題を克服できることから、盛んに研究されてきた。しかし、提案されている多くのハイブリッド手法は､相互作用数のような非ガウス性の変数を扱っているにもかかわらず、ガウス分布に基づく協調フィルタリングを用いて推薦を行うものである。本研究では、交流数をポアソン分布としてモデル化した協調フィルタリングと、協調フィルタリングと潜在変数を共有する変分オートエンコーダーベースのコンテンツ情報生成プロセスに基づくハイブリッド推薦フレームワークを用いる手法を提示している。共有された潜在変数の事前分布として、ポアソン分布の共役事前分布であるガンマ分布を使用する。暗黙的な微分に基づく再パラメータ化のトリックにより、変分オートエンコーダでガンマ分布の利用が可能になる。生成モデルの潜在変数は、確率勾配変分ベイズ法を用いて、ユーザーとアイテムの相互作用の回数とコンテンツ情報を入力として推定する。この推定値に基づき、ユーザーとアイテムの間の観察されない相互作用を推薦のために予測する。コンテンツ情報にはニューラルネットワークベースの生成モデルを用いることで、様々な種類のコンテンツ情報を扱うことができる。実験の結果、提案手法はコンテンツ情報を効果的に利用してインタラクション数を予測することができ、コールドスタート問題の克服に役立つことが確認された。

（２）については、露出バイアスがある場合を考え、二項分布に基づく協調フィルタリングの評価比率を利用したバイアス低減を検討している。観測データにおけるバイアスの扱いは、多くの統計学や機械学習アプリケーションにおける主要な課題の一つである。これまで、推薦システムの文脈では、因果推論に基づく様々な手法が検討されてきた。本研究では、偏った観察から評価を頑健に予測する協調フィルタリング技術について検討している。異なるデータソースにおける偏りのない評価の割合を利用し、協調フィルタリングによる評価の推定割合を偏りのないデータの割合に近づける項を追加することで、協調フィルタリングを拡張する。提案手法は、二項行列分解を用いた協調フィルタリングに基づくものである。偏りのない評価の割合を確率分布として扱い、協調フィルタリングによる推定値との KL ダイバージェンスを最小化する制約項を加えることで、偏りを含む観測から偏りの少ない推定値を得ることを目的としている。二項分布に基づく協調フィルタリングでは、観測値を離散変数として扱うことができるため、偏りのない割合と協調フィルタリングで推定した割合を直接比較することができる。提案手法は、逆傾向重み付け行列分解などの因果推論アプローチによる協調フィルタリング技術の拡張にも適用可能であることを示している。実験結果から、提案手法の単独実装と逆傾向スコア重み付け行列分解との組み合わせにより、レコメンデーションの性能が向上することを示している。

本博士論文は計 93 ページ、英語で執筆され、五つの章で構成されている。第 1 章は推薦システムについて､第 2 章ではその代表的な手法であり､本研究のベースライン手法である協調フィルタリングについて概説されている。第 3 章では、協調フィルタリングと外部コンテンツ情報を組み合わせたハイブリッド推薦法について、特に暗黙のフィードバックについて述べられている。第 4 章では、偏りのない評価の割合から、観測値に偏りが含まれる場合への協調フィルタリングの拡張について議論されている。第 5 章では、本研究の結論と今後の課題について述べられている。

博士論文公聴会では、出願者は予備審査で求められた下記の 7 つの説明項目を含め、上

記博士論文の内容を詳細に説明した。なお、下記の 7 項目については項目 6 を除いて博士論文に説明を追加した。

1. コールドスタート問題が緩和されたかどうか
2. 提案手法（１）と（２）の組合せの可能性
3. ポアソン分布の利用における過分散の問題
4. ガンマ分布の利用とその解釈性
5. これまでに（不採択会議も含む）査読で指摘を受けた重要な論点
6. 平均二乗誤差（MSE）や予測ランキング（NDCG）の評価指標
7. 提案手法（２）における実験パラメータ（協調フィルタリングと KL の重み）

この説明後、参加者より次の項目について質問があったが、出願者は真摯に回答していた。

- 異なるコンテンツ情報の重みで合成したデータに対する MSE、NDCG の違い
- 実データの分布の確認
- MSE、NDCG の有効性

博士論文公聴会後には博士論文審査会が行われた。審査会では、予備審査で要求された項目が博士論文に追加されたか確認するとともに、審査委員の所属など軽微な修正をコメントし、審査委員会は本論文が学位の授与に値すると全員一致で判断した。