情報推薦のための機械学習における学習データのバイアスの可視化

栃木 彩実† 伊藤 貴之† Xiting Wang‡

†お茶の水女子大学大学院 人間文化創成科学研究科 〒112-8610 東京都文京区大塚 2-1-1
‡ Microsoft Research Asia 〒100080 Tower 2, No. 5 Danling Street, Haidian District, Beijing, P.R. China
E-mail: †{g1620527, itot}@is.ocha.ac.jp, ‡xitwan@microsoft.com

あらまし 映画などのコンテンツ推薦システムでは，その推薦エンジンに機械学習を適用することがある．一方で近年，機械学習における公平性やバイアスについての議論が活発化している．学習データのバイアスはこのような問題の一要因であり，不公平な学習結果を引き起こす可能性がある．そこで本研究では学習データと学習結果を比較可視化することで，そのバイアス発見を支援するシステムを提案する．本システムは複数の可視化要素で構成されており，これらの要素を組み合わせて使用することによってデータ分析が可能である．具体例として本報告ではユーザ群の映画鑑賞履歴を学習データとし，機械学習による推薦結果と比較可視化することで，ユーザ間の推薦バイアスの分布を観察した事例を報告する．

キーワード 機械学習，情報推薦，バイアス，公平性，可視化

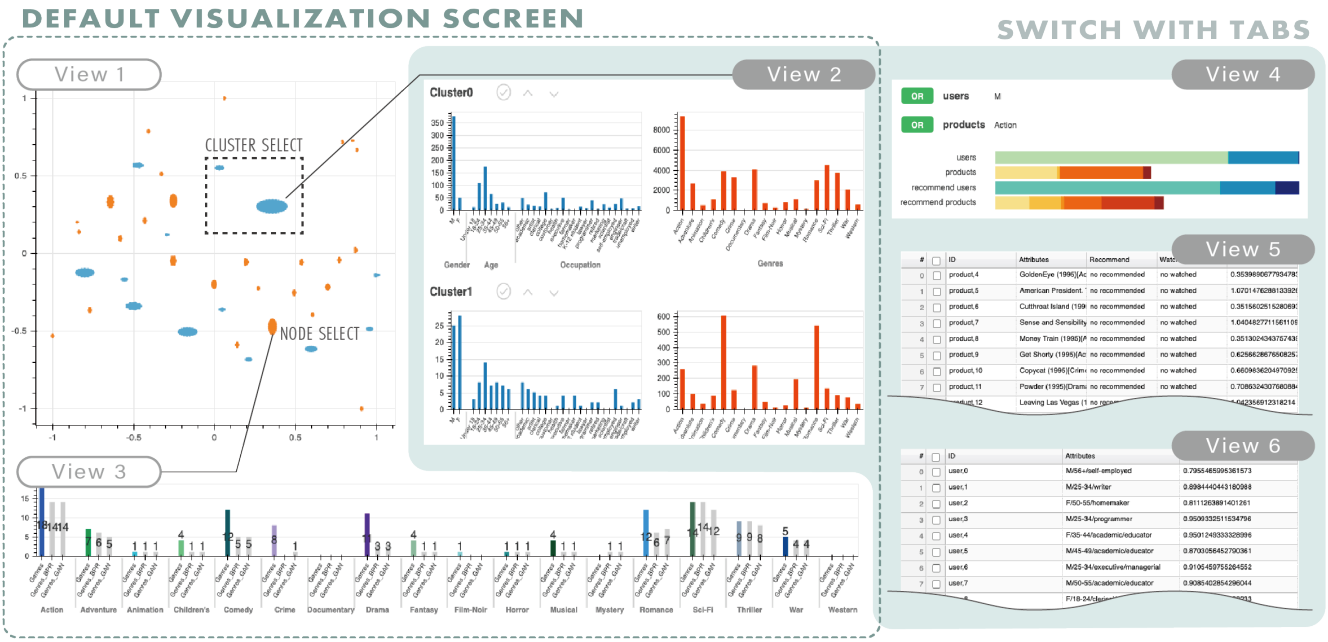

図 1 提案する可視化システムの概要．**View1**: ユーザ群とコンテンツ群のクラスタを示す散布図，**View2**: 各クラスタの属性分布を示す棒グラフ，**View3**: 各ユーザ・コンテンツの属性分布を示す棒グラフ，**View4:** 特定のユーザ・コンテンツ属性の分布を示す帯グラフ，**View5:** 各ユーザ・コンテンツに関する詳細情報を示すデータテーブル，**View6:** 散布図における特定の配色レベルに属するユーザ・コンテンツを表示するデータテーブル．

1. はじめに

インターネットの発達とスマートフォンの普及などによって，人々は場所や時間を問わず買い物や動画・音楽の視聴，SNS での他人との交流などが可能になった．そのようなサービスを提供するシステムは膨大な量のデータを有する．このようなビッグデータを利用する中で，ユーザ自らが好みの商品やコンテンツ，もしくは興味関心のある他ユーザを探し出すのは非常に困難である．そのような背景から，ユーザの代わりにユーザ好みのコンテンツを提案する推薦システムが活躍するようになった．特に，ユーザの過去の行動履歴などといったユーザ固有の情報を分析することによって，各ユーザに対してより適切な推薦が可能である．これによって消費者の購買意欲の促進につながるため，ビジネス面でも推薦システムは大きな影響を与えている．

しかし，過度にユーザ個人に寄り添ってパーソナラ

イズされた推薦システムはユーザにとって喜ばしいものとは限らない．なぜならユーザの嗜好に過度に合わせた推薦結果は，類似コンテンツばかりを提示する可能性が高く，結果的にユーザ自らが知らないうちに触れることのできるコンテンツの幅を狭めてしまうからだ．推薦システムを構築する際にはユーザの嗜好だけでなく，推薦コンテンツの多様性についても考慮する必要がある．

推薦システムには協調フィルタリングをはじめとする多くの手法が知られている．しかしこれらの手法は，過去の行動履歴が少ない新規ユーザへの適切な推薦が難しいというコールドスタート問題が指摘されている．さらに，ユーザやコンテンツといったデータ数の増加にともない,システムへの負荷を高めてしまう．そこで，近年ではこれらの問題を解決するために，推薦システムに搭載されている推薦エンジンに機械学習を適用した事例[1, 2]が増えている．一方で，機械学習には学習モデルや学習データに不公平なバイアスが含まれることで不当な学習結果が得られてしまうという問題がある．特に非常に大規模な学習データを使用した場合，全貌を確認するのは困難である．そのため，高精度の学習モデルを使用していたのにも関わらず，学習データにバイアスが含まれていたために，意図しない不当な結果が得られてしまうという事例が増加している．このような状況を防ぐためには，学習データを分析し精査する必要がある．これは推薦システムに機械学習を適用することを考えた場合でも同様である．前述した過度にパーソナライズされた推薦を防ぐためには，まず学習データ内に偏った嗜好を持つユーザがどう分布しているか調査すべきである．したがって，推薦の多様性や機械学習の公平性といった課題を解決するために学習データの偏りを分析することは非常に重要な手段の 1 つだといえる．

そこで本研究では，学習データと推薦結果を比較可視化することで学習データのバイアス発見を支援する可視化システムを提案する．前述したように，機械学習の学習データのような膨大な量のデータを使用する場合，データを 1 つずつ確認することは非常に困難である．そこでデータを可視化することにより，データ全貌を視覚的に理解することが可能である．さらに，直感的な操作によるデータ探索や分析に要する時間の短縮が期待できる．また，可視化を適用する推薦システムは，ユーザに対して多様なコンテンツを提案するものと想定する．ユーザには年齢・性別・職業が属性として付与され，コンテンツはジャンルで分類されているものとする．

図 1 は提案する本システムの概要について示す．本システムはいくつかの可視化要素から構成されており，これらを組み合わせて使用することでデータ分析を可能とする．まずシステム利用者は **View1** の散布図を用いてユーザ・コンテンツの概要や分布について観察することができる．また **View1** では拡大やノード，クラスタの選択等の機能を備えている．システム利用者は任意のクラスタを選択することで，**View2** 上でそのクラスタにおける属性分布を調査することができる．同様にして，任意のノードを選択した場合は **View3** 上でそのノードにおける属性分布を見ることが可能である．さらにデータ分析を行う中で，**View4** は特定の属性について探索するときに有用であり，**View5, 6** はデータの詳細情報を調べるときに有用である．

2. 関連研究

2-1. 推薦システムのバイアスと多様性

情報推薦に関する研究は長期にわたって数多く発表されている．よく用いられる推薦手法の 1 つに協調フィルタリングがある[3]．Yao ら[4]はこの協調フィルタリングの危険性について言及しており，5 つの公平性指標によって推薦システムの公平性を測る手法を提案している．しかし，全ての指標に対して最適である単一の状況は存在せず，データの特徴や目的に合わせて最も重要な指標を選択しなければならないということも主張している．したがって，推薦システムを構築する上で推薦の公平性を定量的に評価することは大切であるが，そのためには前提としてまずデータの全体像や特徴を知ることが必要であり,重要な課題である．

Farnadi ら[5]は 2 種類の推薦システム固有のバイアスが存在すると述べている．1 つは，同様の推薦事項のみの予測を繰り返すことによって発生する観測バイアスである．このバイアスは，ユーザ自身が無意識に接するコンテンツの幅を狭めてしまう，いわゆる「フィルターバブル」[6]と呼ばれる状況に陥ってしまう要因である．もう一方は，属性や特徴量が偏っている不均衡なデータに起因するバイアスである．このバイアスによる影響は，学習データの属性バランスを均等にした上で再度サンプリングを行った場合でも完全に取り除くことはできなかったと報告されている[7]．

また，情報推薦におけるバイアスを定量化する手法もいくつか発表されている．Beutel ら[8]はペアワイズ比較にもとづいた公平性の評価手法を提案している．Geyik ら[9]は，特定の保護すべき属性の割合を考慮したバイアスの定量化手法を発表している．しかしこれらの手法は，スパース性の高いデータや保護属性が既知でないデータに適応することが困難である．

情報推薦における多様性は，推薦の精度やユーザの満足度を向上するために重要な要素の 1 つである．Vargas ら[10]は，推薦事項の多様化を実現するために

はジャンルの網羅性，冗長性および推薦数について考慮する必要があると主張している．いかに多様なジャンルを推薦するかだけでなく，各ジャンルが推薦される頻度も重要であり，これらは推薦数に依存すると考えられる．その他にもロングテール[11]や目新しさを表すセレンディピティ[12]を考慮することで適切な推薦の多様化が実現できるという研究も発表されている．

2-2. 説明可能な推薦システム

近年，推薦システムに深層学習を適用することによって精度が向上する事例が増えている．しかし，そのような推薦システムでも学習モデルの偏ったもしくは不公平な振る舞いを完全に防ぐことは困難である．そのため，結果として偏った推薦を引き起こしてしまう可能性がある．考えられうる機械学習のバイアスには非常にさまざまな種類が多く存在していることがわかっており[13]，機械学習の公平性を考える上ではそれらに対処する必要がある．それは推薦システムに機械学習を導入する上でも同様である．そこで近年では，前述したような問題の解決を目的とした説明可能な推薦システムに関する研究も発表されている[14, 15].

2-3. 推薦システムと機械学習の可視化

説明可能な推薦システムにとって可視化は非常に強力なツールである.また,機械学習の発展とともに，推薦システムの可視化に関する研究は活発になっている．一方で，推薦システムの公平性やバイアスに焦点を当てた可視化の研究事例はほとんどない．また機械学習の分野では，学習の過程や動作を可視化する研究は多く存在するものの，機械学習の公平性やバイアスを可視化する研究はまだ新しい領域であり，事例もわずかである．

FairVis[16]は人種や性別などのセンシティブな属性を複合的にグループ化し，グループ間で発生する交差バイアスに注目することで，差別や不平等な学習結果を防ぐ可視化解析システムである．各グループの特徴量分布やグループ間の類似度の提示，エントロピーを用いたパフォーマンスの低いグループの検出などを行うことで，ユーザはデータの偏りの発見や，データの部分削除の判断が可能になる．

FairSight[17]は FairDM というフレームワークにもとづく可視化ツールである.FairDM は,データ処理(Data), 学習モデルの選択(Model), 学習結果となるランキングの生成(Outcome)の 3 つのフェーズで構成されている．各フェーズを可視化することで,ユーザは Data で公平性を考慮した特徴量の選択，Model でバイアスの少ない学習モデルの選択，Outcome でランク付けの結果を考慮した公平性の強化をすることが可能である．

これらの事例はどちらもユーザの評価を予測する機械学習に焦点を当てており，ユーザの公平性にのみ着目した可視化システムである．それに対して本論文では，推薦システムに特化した機械学習の公平性に着目し，ユーザと推薦されるコンテンツの双方を可視化する．

3. 提案手法

1 章でも述べたとおり，提案する可視化システムは 6 つの可視化要素で構成されている．可視化するにあたって，以下のようなタスクを定義する．

T1: ユーザおよびコンテンツのパターンの可視化．具体的には類似したコンテンツを評価したユーザのクラスタと，類似したユーザから評価されたコンテンツをクラスタリングし，可視化する．

T2: 不十分な推薦が行われている可能性の高いクラスタの発見.実際の評価と推薦結果に大きな違いがある，もしくは外れ値を含むといったような，偏った推薦を引き起こす恐れのある多様なユーザ・コンテンツのクラスタを発見する．

T3: 特定のユーザ・コンテンツクラスタにおける属性分布の観察.属性分布の特徴を探索することによって，偏った推薦の要因の発見につなげる．

T4: 特定のユーザ・コンテンツにおける属性分布および詳細情報の観察．個々の具体的な推薦結果や属性の特徴を調査することで，偏った推薦の要因についてより深い議論が可能になる．

まずシステム利用者は **View1** を通してデータ全体を観察することにより,**T1** と **T2** を実現する.次に,**View2** を用いて任意のクラスタの特徴を観察することで **T3** が達成される．さらに **View3** で任意のノードの属性分布が表示され，**View5** と **View6** ではユーザ・コンテンツのより詳細な情報が表示される．これらを用いて分析を行うことによって，**T4** の達成につながる．**View4** では，データ分析を行う中で特定の属性について探索したい場合に，システム利用者が単一もしくは複数の属性を選択することでその属性の分布を調査することが可能である．

3-1. 入力データの前処理

本システムで使用するデータは主にユーザと，ユーザに推薦されるコンテンツから構成されている．そこで前処理として,「ユーザノード」と「コンテンツノード」からなる 2 部グラフを構築する．ユーザノードはそのユーザが評価したコンテンツもしくはそのユーザに推薦されたコンテンツをエッジで結ぶものとする．ここで本システムでは視認性低下を防ぐために，初期状態ではエッジを表示しないものとする．次に，接続

関係のあるユーザ・コンテンツの持つ属性の頻度の統計も各ノードに対して算出する．すなわち，ユーザノードの場合は評価したコンテンツの属性頻度f_{ca}と推薦されたコンテンツの属性頻度f_{cr}を計算する．反対に，コンテンツノードでは評価されたユーザの属性頻度f_{ua}と推薦されたユーザの属性頻度f_{ur}を計算する．算出された各属性頻度は特徴量ベクトルとして扱う．また 3-2 節で詳細について言及するが，散布図のノード配色に用いる定量評価値も各ノードに対して事前に算出しておく．

3-2. ユーザ・コンテンツクラスタの散布図

View1 はユーザ群とコンテンツ群をクラスタリングした結果を示す散布図である．クラスタリング手法やノード配置については Koala[19]というグラフ可視化手法を適用している．Koala では，各ノードの特徴量ベクトルや 2 部グラフの接続関係の共通性をもとにクラスタリングを行なっている.また,特定範囲の拡大，クラスタやノードの選択が可能である．さらに以下に示す複数の評価指標を用いたノード配色機能を備えている．

(1) **Similarity**: 散布図上で選択したノードとの類似度にもとづいてノードを配色する．類似度はユーザ間もしくはコンテンツ間のみで算出される．例えばユーザu_iとユーザu_jの類似度は計算する場合は，各ユーザの評価結果における特徴量ベクトル(f_{u_ia}とf_{u_ja})のコサイン類似度を計算することとなる．
(2) **Difference**: 各ノードの評価結果と推薦結果の違いにもとづいてノードを配色する．具体的には，評価結果と推薦結果の特徴量ベクトルのコサイン類似度を算出する．
(3) **nDCG**: ランク付けされた推薦を評価する代表的な指標．本研究では，Järvelin ら[18]の手法を用いて各ノードに対する nDCG を算出．
(4) **Coverage**: 推薦結果に含まれる属性の多様性を評価する指標．データセット全体に存在する全ユーザ・コンテンツ属性に対する，各ノードの推薦結果に含まれる属性数の割合を算出する．
(5) **Weighted Coverage**: **Coverage** モードとは異なり，全ユーザ・コンテンツ属性に対する割合ではなく，各ノードの評価結果に含まれる属性数に対する割合を算出する．すなわち，各ノードの評価傾向に合わせた多様性を評価することが可能である．
(6) **Binomial Diversity**: Vargas ら[10]が定義する網羅性と冗長性から多様性を評価する指標．

各配色モードはユーザとコンテンツそれぞれに対して得られた定量評価値の範囲を 10 段階に分割し，それをもとにノード配色が行われている．**Difference** モードでは類似度が低いほど，すなわち学習データと推薦結果の違いが大きいほど配色の彩度は高くなる．それ以外のモードでは評価値と彩度は比例関係にある．また定量評価値が 0 の場合は，どのノードもグレーで塗られる．

その他にも **View1** の機能として，配色がグラデーションになるような各クラスタ内のノード配置の並べ替えや，エッジの表示の切り替えが可能である．

3-3. 各クラスタにおける属性分布の棒グラフ

View1 上でマウスをドラッグし特定範囲を選択すると，範囲内のクラスタの属性分布が **View2** 上で表示される．ユーザクラスタを選択した場合は，そのクラスタに属するユーザが持つ属性の統計とそれらのユーザが評価したコンテンツの持つ属性の統計がそれぞれ棒グラフとして表示される．同様にしてコンテンツクラスタを選択した場合には，クラスタ内に属するコンテンツの属性分布とそれらのコンテンツを評価したユーザの属性分布が可視化される．ここで，図 1 の **View2** に示すように，水色の棒グラフがユーザ属性の統計，オレンジの棒グラフがコンテンツ属性の統計を表すものとする．

3-4. 各ユーザ・コンテンツにおける属性分布の棒グラフ

View1 上で任意のノードを選択すると，そのノードに関する属性分布が **View3** 上で表示される．ここではユーザ・コンテンツに関する評価結果(f_{ua}, f_{ca})と推薦結果(f_{ur}, f_{cr})を棒グラフとして表示する．こうすることで評価結果と推薦結果を比較して観察することが可能である．本システムでは，推薦結果の棒グラフはグレー，推薦結果の棒グラフはそれ以外の色で塗られている．

3-5. 特定のユーザ・コンテンツ属性における帯グラフ

View4 ではシステム利用者が選択した複数の属性をグループ化し，その属性グループに属するユーザ・コンテンツの割合等を帯グラフとして表示する.ここで，選択する属性は必ずしも複数である必要はなく，単一でもかまわない．しかしここでは選択した属性が単一であっても属性グループと呼ぶことにする．

属性グループは以下の 3 つのパターンが考えられる．

G1: ユーザ属性のみから構成されたグループ
G2: コンテンツ属性のみから構成されたグループ
G3: ユーザ・コンテンツ属性両方を含むグループ

各グループに対して以下のような異なる項目が与えら

れ，それをもとに帯グラフが生成される(図 2).

G1-u1: 該当属性を持つユーザ

G1-u2: 入力データ内の全ユーザ

G1-c1: 該当属性を持つユーザが評価したコンテンツ

G1-c2: 入力データ内の全コンテンツ

G1-c3: テストデータのうち該当属性を持つユーザに推薦されたコンテンツ

G1-c4: テストデータのうち 1 人以上に推薦された全コンテンツ

G1-c5: 全データ(訓練データとテストデータ)のうち，該当属性を持つユーザに推薦されたコンテンツ

G1-c6: 1 人以上に推薦された全コンテンツ

G2-c1: 該当属性を持つコンテンツ

G2-c2: 入力データ内の全コンテンツ

G2-u1: 該当属性を持つコンテンツを評価したユーザ

G2-u2: 入力データ内の全ユーザ

G2-u3: 全ユーザのうち該当属性を持つコンテンツを推薦されたユーザ

G2-u4: 入力データ内の全ユーザ

G3-u1: 該当属性を持つユーザのうち，該当属性を持つコンテンツを評価したユーザ

G3-u2: 該当属性を持つユーザ

G3-u3: 該当属性を持つコンテンツを評価した全ユーザ

G3-u4: 入力データ内の全ユーザ

G3-u5: 該当属性を持つユーザのうち，該当属性を持つコンテンツを推薦されたユーザ

G3-u6: 該当属性を持つコンテンツが推薦された全ユーザ

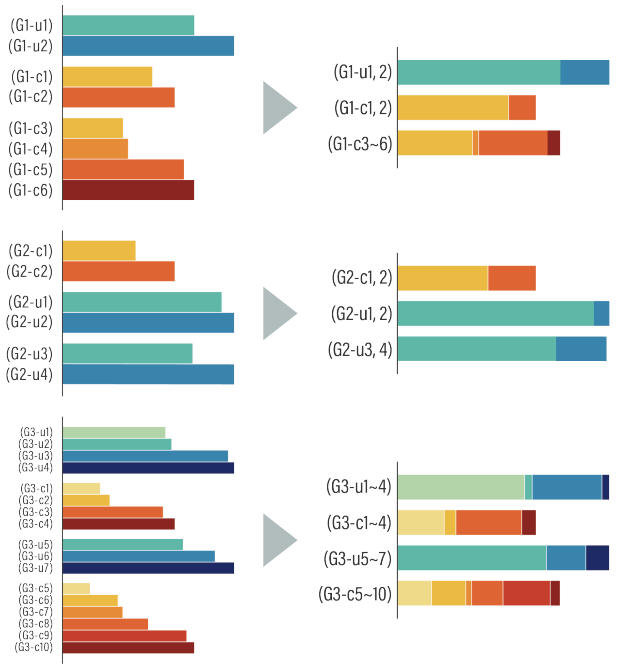

図 2　帯グラフの生成．上から順に **G1**，**G2**，**G3** の帯グラフを表す．

G3-u7: 入力データ内の全ユーザ

G3-c1: 該当属性を持つコンテンツのうち，該当属性を持つユーザに評価されたコンテンツ

G3-c2: 該当属性を持つコンテンツ

G3-c3: 該当属性を持つユーザに評価された全コンテンツ

G3-c4: 入力データ内の全コンテンツ

G3-c5: テストデータ内の該当属性を持つコンテンツのうち，該当属性を持つユーザに推薦されたコンテンツ

G3-c6: テストデータのうち，該当属性を持つユーザに推薦された全コンテンツ

G3-c7: テストデータのうち 1 人以上に推薦された全コンテンツ

G3-c8: 全データ内の該当属性を持つコンテンツのうち，該当属性を持つユーザに推薦されたコンテンツ

G3-c9: 全データのうち該当属性を持つユーザに推薦されたコンテンツ

G3-c10: 1 人以上に推薦された全コンテンツ

また **View4** は **View1** とインタラクティブに機能する．**View4** の帯グラフの各項目を選択すると，その項目内に含まれるユーザ・コンテンツを抽出し，それに応じたユーザ・コンテンツノードをハイライトすることができる．さらに各ユーザ・コンテンツの評価数や推薦数などでフィルタリングすることも可能であり，より絞った条件でノードをハイライトすることができる．

3-6. 詳細情報を記載するデータテーブル

本システムでは **View5** と **View6** にて 2 種類のデータテーブルを表示する．**View5** は **View1** 上で選択したノードに応じた他ノードの詳細情報を示す．このデータテーブルは ID，属性や各ノードの名称，評価および推薦の有無(評価スコア)，ノード間距離といったカラムを持つものとする．例えばユーザノードを選択した場合，そのユーザノードと全コンテンツの評価関係および推薦関係を一覧で表示することができる．また評価や推薦の有無だけでなく，実際の評価スコアもともに表示される．**View6** では **View1** の任意の配色モードに対して，特定の配色レベルであるユーザ・コンテンツを一覧として提示する．このデータテーブルでは ID や属性等の名称に付随して，配色レベルを決定する定量評価値も表示される．また，どちらもデータテーブルも各行を選択することが可能であり，選択した行に対応するユーザ・コンテンツノードが **View1** 上でハイライトされる．

4. 実行結果

4-1. 推薦モデルおよび可視化データの構築

本研究では映画鑑賞データセットを用いた．これは 3883 本の映画と 6040 人の鑑賞者を含み，映画はジャンル，鑑賞者は性別，年齢，職業の属性を持つ．また学習モデルとして，BPR[1]によるモデル，BPR に GAN[2]を適用したモデルの 2 種類を使用する．本研究はデータセットを訓練データとテストデータに分割した上で，訓練データに対して機械学習を適用する．この機械学習の学習結果にもとづき，テストデータに含まれる各鑑賞者に対して上位 20 本の映画を推薦した．そして鑑賞者 1000 人とその鑑賞者らが視聴した映画 512 本をサンプリングして可視化した．

4-2. 可視化画面の観察

上記のデータセットを可視化した一例を紹介する．まず **View1** を **Difference** モードで配色し，各クラスタを観察すると，図 3-1，3-2 の散布図に示すように，彩度が低いノードが多いユーザクラスタ 2，10 と彩度が高いノードが多いクラスタ 1，5 を確認できる．視聴結果と推薦結果の差の大きさと彩度は比例する．すなわち，彩度が低いクラスタは推薦結果との差が小さいユーザが多く，彩度が高いクラスタは推薦結果との差が大きいユーザが多いと考えられる．そこで，**View2** を用いて各クラスタの属性分布を観察する．図 3-1 のクラスタ 2，10 のユーザ属性の統計を示す棒グラフを見ると，どちらも女性に比べて男性が非常に多いことがわかる．一方で，図 3-2 のクラスタ 1，5 を見るとどちらもクラスタ 2，10 に比べて女性の割合が高いことが見てとれる．さらに，クラスタ 5 では年齢層も他のクラスタより若く，子供が多いクラスタであることがわかる．ここで，本研究で用いたデータセットのうち，女性や子供といったユーザは少数である．すなわち，少数派ユーザは多数派ユーザよりも適切な推薦が行われにくいといえる．実際に推薦結果との差が大きいユーザの例を図 4 に示す．上記クラスタから 1 人のユーザを選択し，**View3** 上で視聴結果と推薦結果を比較した．図 5 を見ると，Children's や Animation を多く視聴しているのに対して Comedy や Sci-Fi が多く推薦されていることがわかる．これは当該ユーザの視聴傾向が大衆と異なることが要因と考えられる．その他にも視聴数が非常に少ないユーザは推薦結果が適切でない場合が多く見られた．

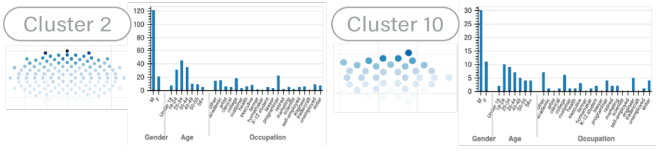

図 3-1 視聴結果と推薦結果との差が小さいユーザが多いクラスタにおける **View1** と **View2** の可視化結果

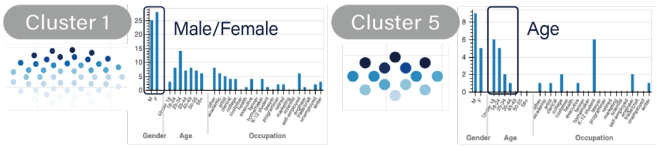

図 3-2 視聴結果と推薦結果との差が大きいユーザが多いクラスタにおける **View1** と **View2** の可視化結果

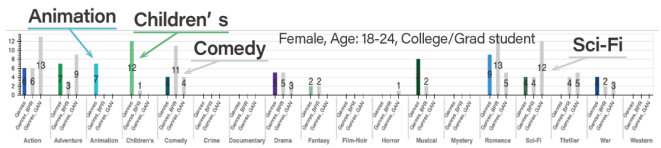

図 4 視聴結果と推薦結果の差が大きいユーザにおける **View3** の可視化結果

次に男性ユーザと女性ユーザにおける推薦の違いの要因についてさらに掘り下げていくこととする．図 5 は前述したクラスタ 1 と 2 に属するユーザがそれぞれ視聴したコンテンツ属性の統計の棒グラフである．このグラフを見ると，男性が多いクラスタ 2 は Action や Sci-Fi を多く視聴しており，女性が多いクラスタ 1 では Comedy や Romance を視聴していることが確認できる．そこで本報告では，男性は Action，女性は Romance といった特定のジャンルが強く影響しており，それらが推薦結果の偏りの要因であるという仮説を立てた．

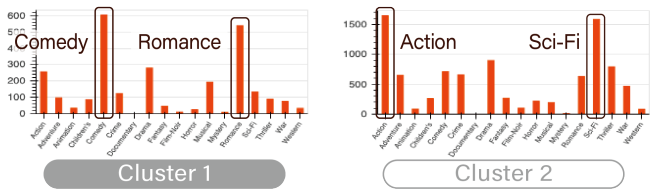

図 5 各クラスタのコンテンツ属性の統計

探索する属性を性別と Action，Romance の 2 種類のジャンルに限定する．**View4** 上で性別とジャンルの計 4 種類の組み合わせの属性グループを作成し，図 6 に示すような帯グラフを生成した．各帯グラフの視聴ジャンルの割合に着目すると，男性ユーザは Action の視聴割合と Romance の視聴割合にさほど大きな違いは見られない．一方で，女性ユーザは Romance の視聴割合に比べ Action の視聴割合に大きな差が確認できる．また **View4** で Action と Romance それぞれの帯グラフを生成し **G2-c1** を選択することで，各ジャンルのコンテンツクラスタの分布を散布図上で可視化する．図 7 を観察すると，Action は広範囲に分布しているのに対し Romance は一部分に固まって分布していることがわかる．これは Action の方が他のジャンルとの組み合わせ

が多いことや，幅広いユーザに視聴されている可能性が考えられる．このことから，女性と Romance が男性と Action に比べ，より強い偏りを持った視聴傾向があるのではないかと予測できる．

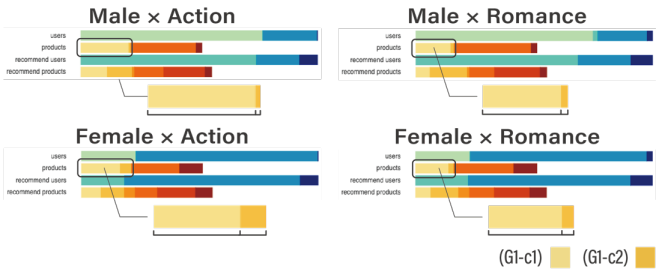

図 6　性別とジャンルからなる帯グラフ

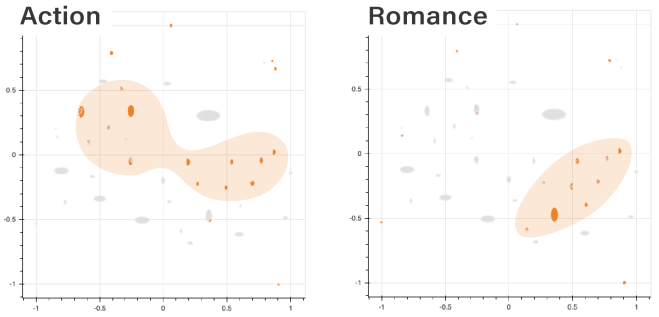

図 7　Action，Romance 映画の散布図上の分布

そこで視聴数でフィルタリングすることによって，より限定したユーザを対象に可視化を行なった．図 6 で示した 4 つの帯グラフの **G3-u1** を選択し，**View1** 上で該当するユーザをハイライトする．本報告では視聴数が 20 本以上のユーザのみハイライトされるように，フィルタリングを施している．ここで，本システムではハイライトされていないノードは全てグレーで塗られるよう設定されている．

図 8 は Action と Romance の視聴数がそれぞれ 20 本以上であるユーザの散布図の結果を示す．散布図は **Difference** モードで配色が行われている．図 8 の線で囲われたノードは両ジャンルとも 20 本以上視聴したユーザを表す．また該当する男性ユーザの多くがクラスタ 0 とクラスタ 6 に集中し，該当する女性ユーザの多くがクラスタ 3 に集中したため，これら 3 つのクラスタに注目して観察する．まずクラスタ 0，6 の Action と Romance それぞれの結果を比較すると，Romance を多く視聴する男性ユーザのほとんどが Action を多く視聴していることがわかる．これらのユーザの視聴傾向を **View3** にて観察すると，図 9-1 の例のようにジャンルを問わず幅広く多様なジャンルを視聴しているユーザであることが読み取れる．また図 9-2 を見ると，Action のみを 20 本以上視聴したユーザの場合でもある程度多くのジャンルをバランスよく視聴していることが確認できる．前述したように Action は汎用性の高いジャンルであり，他の多様なジャンルと共存しやすい．したがって，このようなユーザは Action を好んで視聴しているものの，結果として Action と組み合わされた多様なジャンルを視聴していると推測できる．

一方で女性ユーザの場合は男性ユーザとは異なる結果となった．図 8 のクラスタ 3 における Action と Romance それぞれの散布図を比較すると，Action を多く視聴する女性ユーザのほとんどが Romance も多く視聴していることがわかる．これは男性の場合と同様の傾向が見られた．図 10-1 の例に示すように，Action や

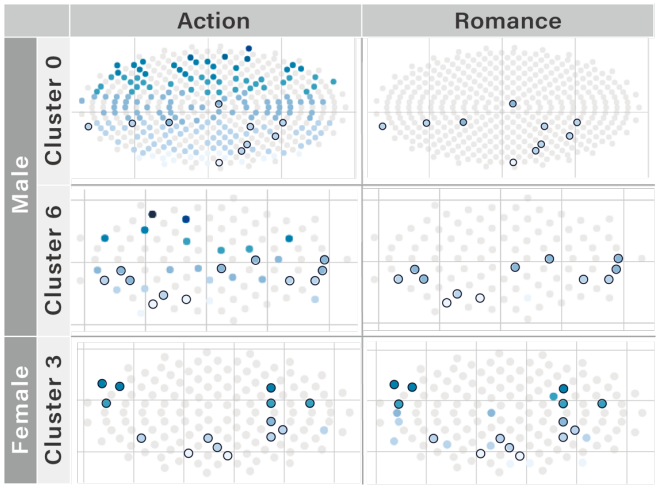

図 8　Action，Romance を 20 本以上視聴したユーザの散布図の可視化結果

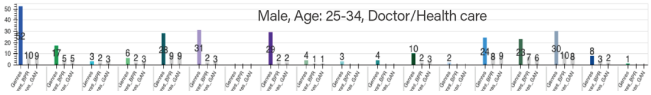

図 9-1　Action と Romance をともに 20 本以上視聴した男性ユーザ

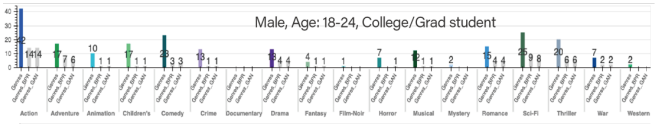

図 9-2　Action のみ 20 本以上視聴した男性ユーザ

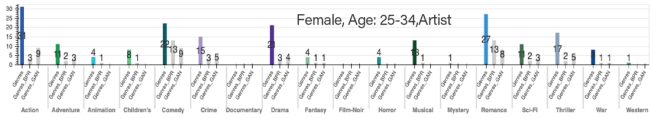

図 10-1　Action と Romance をともに 20 本以上視聴した女性ユーザ

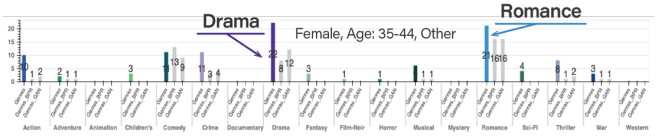

図 10-2　Action のみ 20 本以上視聴した女性ユーザ

Romance に限らず多様なジャンルを視聴しているユーザであることが確認できる．しかし，Romance のみを 20 本以上視聴した女性ユーザに着目すると，視聴ジャンルが偏っているユーザが多いことがわかった．図 10-2 は Romance のみを多く視聴する女性ユーザの視聴傾向の一例である．この例で示すように，Romance や Drama などの視聴数が顕著である．このことから，Romance を多く視聴する女性ユーザは偏った視聴傾向である場合が多いといえる．したがって，本研究で使用したデータセットには,「Romance は女性が好んで視聴するジャンルである」というバイアスがかかっている可能性が高いことが示唆される．

5. まとめと今後の展望

本報告では推薦システムの公平性に着目し，複数の可視化要素を組み合わせて学習データと推薦結果を比較可視化することで，学習データに潜むバイアスの発見を支援するシステムを提案した．

今後の展望として本システムの拡張があげられる．機械学習が外れやすいクラスタにはどのような特徴があるのかを分析し，システムの利用者に対して学習データから外すべきノードやクラスタなどを強調表示したいと考えている．この機能を実現することによって，システム利用者がより直感的に学習データを分析し，精査することが可能になる．

参 考 文 献

[1] S. Rendle, C. Freudenthaler, Z. Gantner and L. Schmidt-Thieme, “BPR: Bayesian Personalized Ranking from Implicit Feedback, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence”, pp. 452-261, 2009.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets”, Proceedings of Neural Information Processing Systems (NIPS), pp. 2672-2680, 2014.

[3] A. Das, M. Datar, A. Garg and S. Rajaram, “Google News Personalization: Scalable Online Collaborative Filtering”, The Web Conference (WWW), pp. 271-280, 2007.

[4] S. Yao and B. Huang, “New Fairness Metrics for Recommendation that Embrace Differences”, In Workshop on Fairness, Accountability and Transparency in Machine Learning (FAT/ML), arXiv: 1706.09838, 2017.

[5] G. Farnadi, P. Kouki, S. K. Thompson, S. Srinivasan and L. Getoor, “A Fairness-aware Hybrid Recommender System”, In the 2nd FATREC Workshop on Responsible Recommendation, arXiv: 1809.09030, 2018.

[6] E. Pariser, “The Filter Bubble: What the Interest Is Hiding from You”, Penguin Press, 2011.

[7] M. D. Ekstrand, M. Tian, I. Madrazo Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill and M. Soledad Pera, “All The Cool Kids, How Do They Fit In? Popularity and Demographic Biases in Recommender Evaluation and Effectiveness”, Machine Learning Research, Vol. 81, pp. 1-15, 2018.

[8] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi and C. Goodrow, “Fairness in Recommendation Ranking through Pairwise Comparisons”, ACM SIGKDD Conference on knowledge discovery and Data Mining, pp. 2212-2220, 2019.

[9] S. C. Geyik, S. Ambler and K. Kenthapadi, “Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search”, ACM SIGKDD Conference on knowledge discovery and Data Mining, pp. 2221-2231, 2019.

[10] S. Vargas, L. Baltrunas, A. Karatzoglou and P. Castells, “Coverage, redundancy and size-awareness in genre diversity for recommender systems”, ACM Conference on Recommender systems, pp. 209-216, 2014.

[11] L. Shi, “Trading-off Among Accuracy, Similarity, Diversity, and Long-tail: A Graph-based Recommendation Approach”, ACM Conference on Recommender system, pp. 57-64, 2013.

[12] M. Kaminskas and D. Bridge, “Measuring Surprise in Recommender Systems”, In Proceedings of the Workshop on Recommender Systems Evaluation: Dimensions and Design, pp. 393-394, 2014.

[13] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, “A Survey on Bias and Fairness in Machine Learning”, ACM Computing Surveys, Vol. 54, No. 6, pp 1-35, 2021.

[14] X. Wang, Y. Chen, J. Yang, L. Wu, Z. Wu and X. Xie, “A Reinforcement Learning Framework for Explainable Recommendation”, IEEE International Conference on Data Mining, 2018.

[15] J. Gao, X. Wang, Y. Wang and X. Xie, “Explainable Recommendation Through Attentive Multi-View Learning”, AAAI Conference on Artificial Intelligence, pp. 3622-2629, 2019.

[16] A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern and D. H. Chau, “FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning”, IEEE Conference on Visual Analytics Science and Technology, arXiv: 1904.05419, 2019.

[17] Y. Ahn and Y. Lin, “FairSight: Visual Analytics for Fairness in Decision Making”, IEEE Transactions on Visualization and Computer Graphics, Vol. 26, No. 1, pp. 1086-1095, 2020.

[18] K. Järvelin and J. Kekäläinen, “Cumulated gain-based evaluation of IR techniques, ACM Transactions on Information Systems”, Vol. 20, No. 4, pp. 422-446, 2002.

[19] T. Itoh and K. Klein, “Key-node-Separated Graph Clustering and Layout for Human Relationship Graph Visualization”, IEEE Computer Graphics and Applications, Vol. 35, No. 6, pp. 30-40, 2015.