情報推薦のための機械学習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近年，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．このような問題の一要因として学習データのバイアスが挙げられる．学習データにバイアスが存在することで，意図せずとも不公平な学習結果を生じることがある．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，そのバイアスの発見につなげるシステムを提案する．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，機械学習による映画推薦結果と比較可視化することで，ユーザ間の推薦のバイアスがどのように分布するか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 (1)選択されたノードに対する特徴量の分布を描画した棒グラフ，(2) ユーザ群とアイテム群を同一座標平面上に描画した散布図，(3)選択されたクラスタに対する特徴量の分布を描画した棒グラフ，(4)選択されたノードに関する詳細な情報を記載したデータテーブル

1. はじめに

インターネットの普及が進み、人々はどんな時間や場所でも，スマートフォンや PC などのデバイスさえあれば，買い物や動画・音楽の視聴，SNS による他人との交流が可能になった．それらを使用するサービスは膨大な量の情報を所持する．したがって，その中からユーザ自らが好みの商品や映像，音楽などのコンテンツ，もしくは，共通の関心を有する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で，各ユーザに対してより適切な推薦が可能になる．これによって，消費者の購買意欲を引き出すことにつながり，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，ユーザ個人に寄り添ったパーソナライズな推薦システムが普及することによって，ユーザの知らないうちに，限られた範囲の情報やコンテンツのみが提示され，新たな話題やジャンル，視点などに出会う機会を奪われることが懸念される．このような状況を防ぐためにも，推薦システムを構築するには公平性や多様性を考慮しなければならないという課題が挙げられる．

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には，商品数やユーザ数の増加にともなうサービスの大規模化がシステムの負荷を高めるという問題や，初期ユーザへの適切な推薦の難しさを指摘するコールドスタート問題などが知られている．そこで近年では，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する事例[1, 2]が増えている．しかし，機械学習は不公平なバイアスを含む学習データや学習モデルを利用することで，不当な学習結果を引き起こす可能性がある．近年このような機械学習におけるバイアスや公平性について問題視されている．特に，非常に大規模な学習データの全貌を確認するのは困難である場合が多い．そのため，実際に高精度の学習モデルを使用したにもかかわらず，学習データにバイアスが含まれていたために，意図せず不当な学習結果が得られてしまう事例が増えている．このような状態を防ぐためにも，学習データを精査することは重要である．

そこで，本報告は推薦システムにおける機械学習の学習データに着目し，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視線する可視化システムを提案する．本システムは，以下の可視化手法で描画されたパーツを組み合わせることによって，図 1 で示すような，1 つの可視化画面を構築している．

- ユーザ群と推薦されるアイテム群をノードとして各々にクラスタリングを適用した結果を画面配置した散布図.
- 散布図上の各ノードにおけるユーザまたはアイテムの特徴量を表す棒グラフ.
- 散布図上の各クラスタの特徴量を表す棒グラフ.
- 散布図上の各ノードにおけるデータの詳細を記載したデータテーブル.

散布図では，各ノードの特徴量から類似度を算出し，その分布をノードの色で表現する．また，各ノードにおける棒グラフでは学習データと推薦結果の値を比較することが可能であり，学習データと推薦結果に大きな差異があるかを視覚的に観察することができる．

本報告の構成は以下のとおりである．2 章では関連研究について述べる．3 章では提案手法について，4 章では本手法の実行結果と実際に観察した事例について述べる．5 章では本研究のまとめと今後の課題，展望について述べる．

2. 関連研究

2-1. 推薦システムにおける公平性

推薦システムによく用いられるアルゴリズムの一つとして協調フィルタリングがあげられる．協調フィルタリングは観測されたデータに基づく予測手法である．したがって推薦結果は，データ内に存在するバイアスの影響を受ける．言い換えれば，偏ったデータは少数派のユーザグループに対して不公平な予測結果を導く可能性が高いことを述べている．Yao ら[3]は，こういった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な状況は存在せず，データの特徴や目的に合わせて最も重要な公平性の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．そのためにはデータの特徴や全体像を知ることが重要であり,推薦システムを構築する上での課題となっている．

また，Farnadi[4]らは，推薦システムには 2 種類の固有バイアスが存在すると述べている．一つは，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．もう一方は，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．これ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」[5]と呼ばれる状況に陥る要因である．さらに, Leonhardt ら[6]は，推薦システムにおける公平性と多様性はトレードオフの関係であると主張している．ここで述べられている多様性とは，全ユーザに提示されるコンテンツの偏りをなくすことを意味する.推薦システムの有用性，すなわちユーザの好みに則した予測結果を鑑みた上で多様性を考慮する場合，例えば購入額が大きいユーザには高評価のコンテンツばかり推薦され，購入額が少ないユーザには品質の判断が難しい新しいコンテンツばかり推薦されるといった事態が生じると述べている．本研究では，以上で言及されているような多様なユーザに対して適切なコンテンツを推薦できているかを，提案する可視化システムを通して観察する．ここでの多様なユーザとは例えば，特定の属性や特徴を持つコンテンツばかりを選択しているユーザや，そもそも観測データが少ないユーザなどを意味する．

2-2.機械学習の可視化

機械学習の発展とともに明らかになってきたブラ

ックボックス化といった課題を，可視化で解決する研究[7, 8]が近年多く発表されている．これらの研究は機械学習の透明性を保持する目的や，機械学習の仕組みを学ぶための支援ツールとしての活用が期待されている．一方で，機械学習の公平性に着目した可視化システムに関する研究はまだ新しい領域であり，事例もわずかである．以下，その事例を簡単に紹介する．

FairVis[9]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[10]は，FairDM というフレームワークに基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．それに対して本研究では，推薦システムに特化して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，現時点での我々の実装では，推薦システムにおける学習データと推薦結果を可視化するシステムを，4 つの可視化画面で構築している．本章では可視化手法とその詳細について述べる．

3-1. データ全体の可視化

本研究では，ユーザ群からなるデータと，推薦されるアイテム群からなるデータの 2 種類に対して機械学習を適用することを想定する．これらの 2 種類のデータ全体を可視化するために提案手法では，両データを統合したデータに対してクラスタリングを適用し，同一の散布図上にユーザ群とアイテム群をノードとして描画する．ここで，クラスタリング手法やノード配置については Koala[11]というグラフ可視化手法を適用している．

また，ノードの配色には以下の 2 つのモードを実装している．

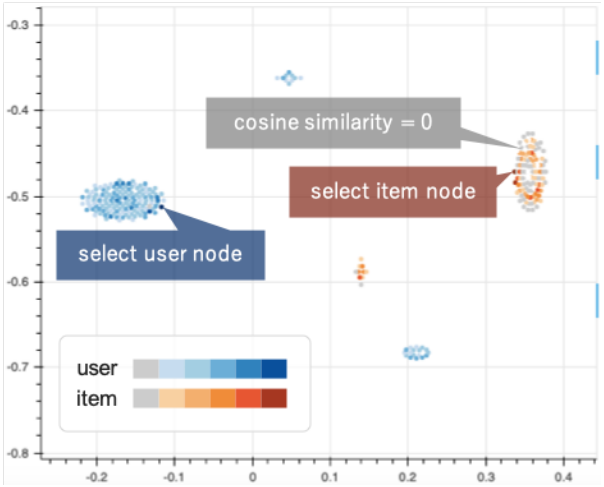

図 2　散布図 (Similarity モード)

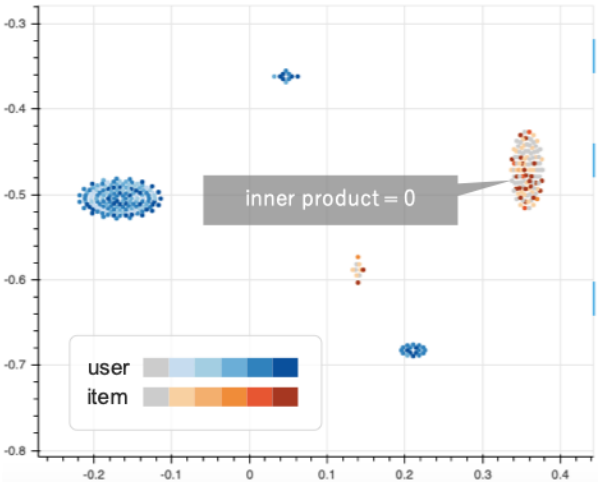

図 3　散布図 (Difference モード)

(1) Similarity: 散布図上のノードを選択すると，選択されたノードとのコサイン類似度にもとづいて他のノードを 5 段階に配色する．類似度が高いノードほど彩度が高くなり，コサイン類似度が 0 である場合にはノードはグレーで描画される．また，選択されたノードは他のノードよりも彩度の高い色で描画される．この配色による描画例を図 1 に示す。

(2) Difference: 学習データの特徴量ベクトルと推薦結果の特徴量ベクトルの内積を各ノードに対して算出し，その値にもとづいて配色を 5 段階に変化させる．内積が小さい，すなわち学習データと推薦結果との差が大きいノードほど彩度が高くなる．ここで，例えば誰にも推薦されなかったアイテムなど，内積が 0 となるノードの色はグレーで描画される．この配色による描画例を図 2 に示す。

ノードの色分布を観察することで，選択したノード

に類似度の高いノードの発見や，推薦結果の適切さの確認を支援できる．

3-2. 各ノードの特徴量の分布

散布図上のノードを選択した際に，そのノードに関連するデータ特徴量の分布を棒グラフで描画する．ここで述べている「関連するデータ特徴量」とは，ユーザノードの場合，ユーザが選択したアイテムの特徴量であり，反対にアイテムノードの場合は，そのアイテムを選択したユーザの特徴量を表す．さらに，ユーザノードを選択した場合は，学習データ(すなわち既にユーザが選択したアイテム)の分布だけでなく，学習データにもとづいて推薦された結果の分布を同時に可視化することが可能である．これによって，選択したノードに対して，学習データと推薦結果を詳細に比較できる．

3-3. 各クラスタ内の特徴量の分布

我々が実装している可視化システムでは，散布図上で 1 つのノードを選択するだけでなく，散布図上の任意の矩形領域をドラッグすることで，散布図上の任意の範囲に含まれる複数のノードを同時に選択することが可能である．これを利用して，選択された矩形領域に包括される各クラスタに関連するデータ特徴量の分布を棒グラフで描画する．ここで各クラスタに対して，ユーザ特徴量とアイテム特徴量の 2 種類の棒グラフを表示する．例えばユーザクラスタを選択した場合，ユーザ特徴量グラフでは，そのクラスタに含まれるユーザ群の特徴量の分布，すなわちどのようなユーザが多く集まっているのかを示す．アイテム特徴量グラフでは，そのクラスタに含まれるユーザ群が選択したアイテムの特徴量の分布，すなわちそのクラスタのユーザはどのような傾向でアイテムを選択しているかを示す．クラスタごとの分布を描画することによって，データ全体の偏りとなるクラスタや，外れ値などの発見につながる．

3-4. 各ノードに関連する詳細情報

散布図上のノードを選択した際に，そのノードに関するデータの詳細をデータテーブルとして提示する．具体的には，ID，特徴量(名前，年齢，職業など)，ユーザの選択の有無，評価値，他ノードとの距離を列挙する．3-2 章の棒グラフが，選択されたアイテムの概略的な特徴を表現しているのに対して，データテーブルを併用することで，アイテムに対する評価結果や，具体的なユーザの選択内容などを観察することができる．なお，データテーブルは列ごとのソートが可能である．

4. 実行結果

4-1. データの前処理

本研究では学習データとして，ユーザ群の映画鑑賞履歴データ[12]を使用する．このデータは，6040 人のユーザと 3883 件の映画からなっており，ユーザと映画それぞれの視聴関係と，1~5 の 5 段階評価結果が含まれる．また，可視化画面では学習データだけでなく，機械学習結果にもとづいて得られた推薦結果も描画する．本報告では，学習モデルとして，

- BPR(Bayesian Personalized Ranking)[1]によるモデル
- BPR と GAN(Generative Adversarial Network)[2]を組み合わせたモデル

の 2 通りを使用する．本研究ではデータセットの中から，ユーザと映画の視聴関係のみを抽出し，これ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．この機械学習結果にもとづいて，テストデータに含まれる各ユーザに対して上位 20 件の映画の推薦結果を求める．本報告では，テストデータの中から，ユーザ 1000 人とそのユーザ群が視聴した映画 512 件を学習データとして可視化画面上に描画している． また，ここではユーザの特徴量として年齢，性別，職業を使用し，アイテムの特徴量としてジャンルを使用する．

4-2. 可視化画面の観察

4-1 章で述べたデータに対して，本研究の可視化システムを利用して観察した結果を示す．

図 4 は，映画鑑賞データを difference モードで描画した散布図である．まず，暖色で塗られたアイテムノードのクラスタに着目すると，グレーで塗られたノードが多いことがわかる．グレーということは，学習デ

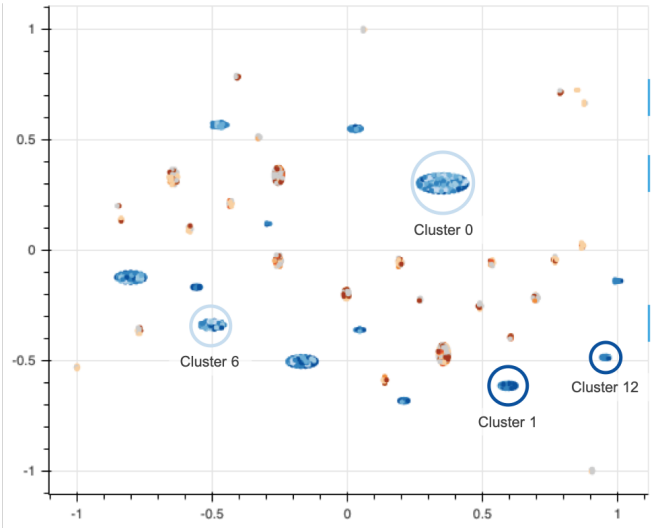

図 4　映画鑑賞者のデータを散布図(difference モード)で描画した例

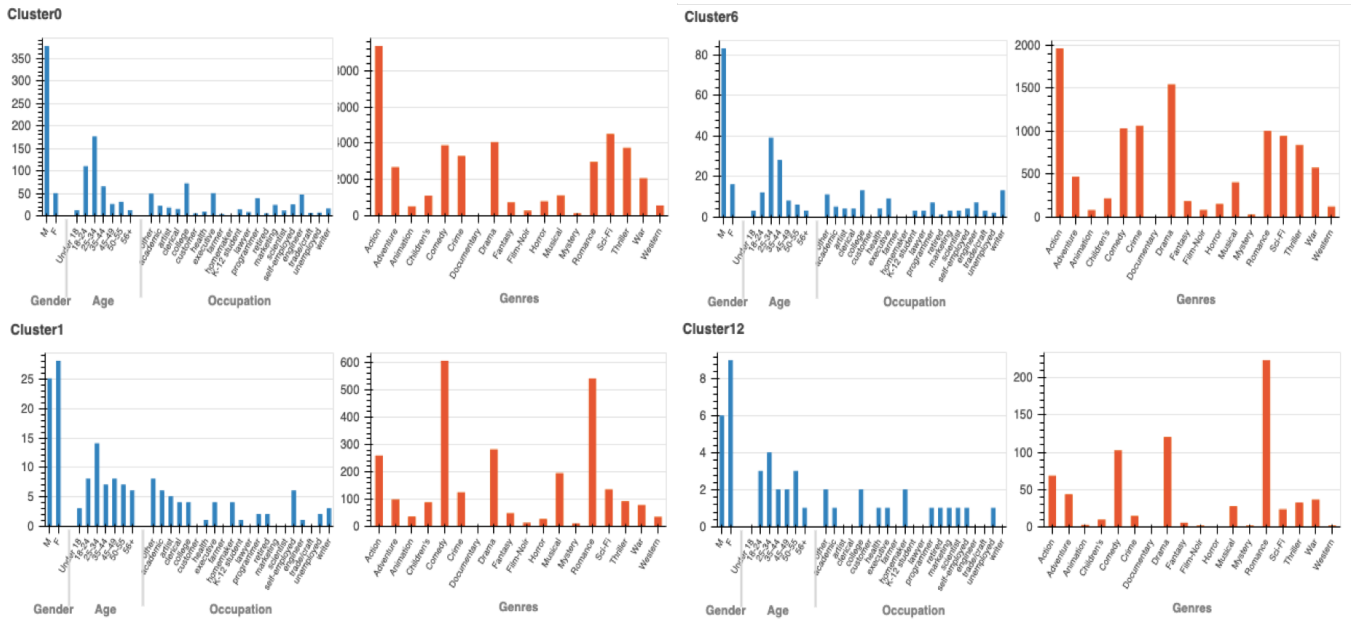

図 5　各ユーザクラスタにおけるユーザ特徴量の分布と
ユーザが選択したアイテム(試聴した映画)のジャンルの分布を描画した棒グラフ

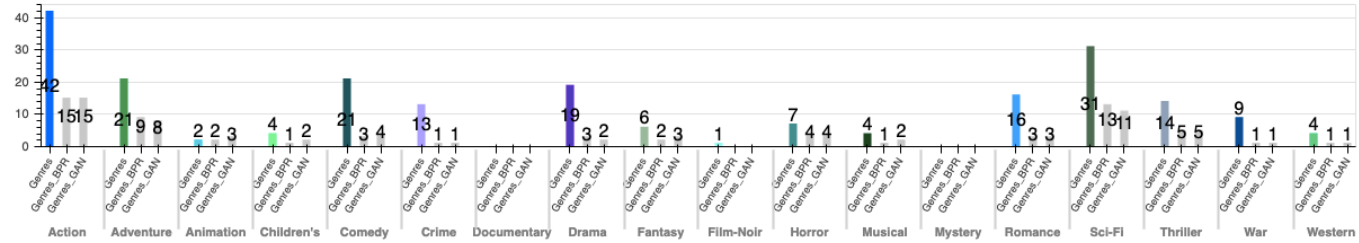

図 6　学習データと推薦結果の分布の差が小さいユーザノードの例

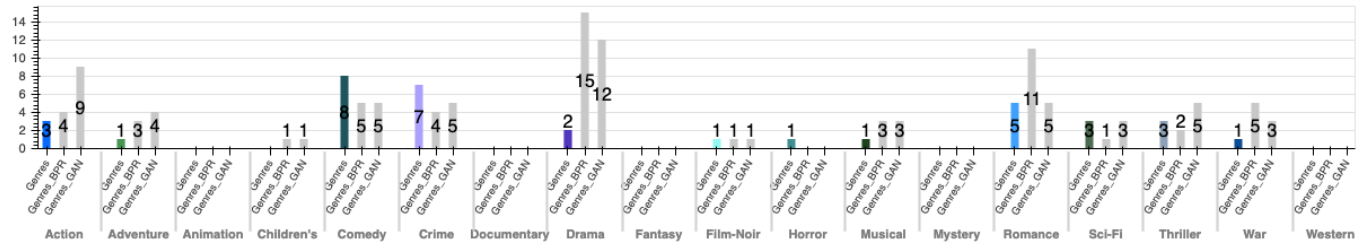

図 7　学習データと推薦結果の分布の差が大きいユーザノードの例

ータの特徴ベクトルと推薦結果の特徴ベクトルの内積が 0 であることを表している．ここで，どのユーザも選択していない，すなわち視聴数 0 のアイテムノードは描画の際に除いている．したがって，多くの場合，グレーのアイテムノードは，どのユーザにも推薦されなかったアイテムである．このことから，本研究で使用した推薦手法では，推薦されるアイテムに偏りがある可能性があると言える．ここで，図 4 では BPR と GAN を組み合わせた手法の場合を描画しているが，BPR 単体の学習モデルと比較しても大きく可視化結果が変化することはなかった．

次に，図 4 中の寒色で塗られたユーザノードに注目する．ユーザ群から形成されるクラスタのうち，他クラスタに比べ，彩度の高いノードが多いクラスタと彩度の低いノードが多いクラスタの特徴を比較する．図 5 より，彩度の低い cluster0 と cluster6 は男性が多く，Action 映画を多く試聴している傾向にある．彩度が高いクラスタ 1 とクラスタ 12 は女性が多く，Romance 映画を多く試聴している傾向があることが読み取れる．このことから，女性の多いクラスタの方が推薦結果と学習結果の差が大きく，推薦結果がユーザの嗜好にそぐわない結果になっている可能性が高い．他のクラス

タを分析しても,男性の割合が大きいクラスタが多く,学習データ全体が男性に偏っていることがわかる.

また，図 6 と図 7 はユーザが試聴した映画のジャンルの分布と,推薦された映画のジャンルの分布を示す.図 6 では試聴した映画数が多いユーザであり，試聴したジャンルの傾向に沿った推薦が実現されていることが読み取れる．一方で，図 7 は試聴した映画数が比較的少ないが，Comedy や Crime といったジャンルの映画を好んで見ていることがうかがえる．しかし，推薦された結果では，BPR 単体と BPR+GAN のどちらも学習モデルを用いても Drama が多く推薦されている．これは，明らかに特定のジャンルを好むユーザであっても，試聴数が少ないと適切な推薦をすることが困難である可能性が示唆される.

5. まとめと今後の課題

本報告では，推薦システムの公平性に着目し，複数の可視化画面を組み合わせて学習データと推薦結果を比較することで，学習データに存在するバイアスの発見を支援するシステムを提案した.

今後の課題として，まず学習データと推薦結果の差異を評価する手法の検討が挙げられる．現在はそれぞれの特徴量ベクトルの内積によって求めている．しかし，アイテムの公平性を考慮した場合，ユーザの好みに沿った推薦が必ずしも公平であるかどうかは限らない．実際に本報告では，散布図の可視化結果から推薦されるアイテムに偏りがあることがわかっており，多様性も考慮した推薦を実現する場合は，学習データと推薦結果の差異以外の部分も考慮する必要がある.

また，本システムの拡張として，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，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-Theime, “BPR: Bayesian Personalized Ranking from Implicit Feedback”, In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 452-461, 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”, In Proceedings of Neural Information Processing Systems (NIPS), pp. 2672-2680, 2014.

[3] 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.

[4] 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.

[5] P. Eli, “The Filter Bubble: What the Interest Is Hiding from You”, Penguin Press, 2011.

[6] J. Leonhardt, A. Anand and M. Khosla, “User Fairness in Recommender Systems”, In Companion Proceedings of the The Web Conference (WWW), pp.101-102, 2018.

[7] M. Kahng, P. Y. Andrews, A. Kalro and D. H. Chau, “ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models”, IEEE Transactions on Visualization and Computer Graphics, Vol.24, No.1, pp.88-97, 2018.

[8] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viègas and J. Wilson, “The What-If Tool: Interactive Probing of Machine Learning Models”, IEEE Transactions on Visualization and Computer Graphics, Vol. 26, No. 1, 2020.

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

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

[11] T. Itoh, 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.

[12] F. M. Harper and J. A. Konstan. “The Movielens datasets: History and context”, ACM Transactions on Interactive Intelligent Systems (TiiS), Vol. 5, No. 4, pp. 19, 2016.