

情報推薦のための機械学習のバイアスの可視化

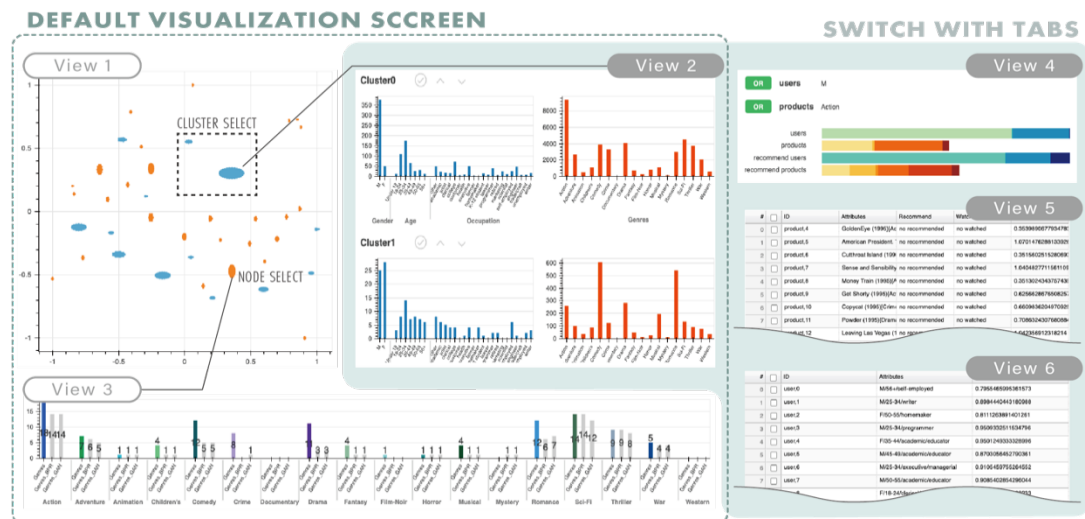
栃木 彩実[†] 伊藤 貴之[†] Xiting Wang[‡]お茶の水女子大学 人間文化創成科学研究科 〒112-8610 東京都文京区大塚 2-1-1[†]Microsoft Research Asia 〒100080 Tower 2, No.5 Danling Street, Haidian District, Beijing, P.R. China[‡]

図 1 提案する可視化システムの概要. **View 1:** ユーザ群とコンテンツ群のクラスターを示す散布図, **View 2:** 各クラスターの属性分布を示す棒グラフ, **View 3:** 各ユーザ・コンテンツの属性分布を示す棒グラフ, **View 4:** 特定のユーザ・コンテンツ属性の分布を示す帯グラフ, **View 5:** 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, **View 6:** 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. 概要

推薦エンジンへの機械学習の適用事例が近年増加している. 一方で, 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おり, 重大な課題の1つとなっている. 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あげられる. 学習データにバイアスが存在することで, 意図しない不公平な学習結果を生じることがある. そこで本報告では, 学習データと学習結果を比較可視化し, バイアス発見につなげるシステムを提案する. 本報告では, 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, ユーザ間の推薦バイアスの分布を観察した事例を報告する.

2. 関連研究

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, 機械学習の公平性やバイアスの可視化は新しい課題であり, 研究事例もまだ少ない[1]. これらの研究では, ユーザ評価を予測する機械学習におけるユーザの公平性のみに着目した可視化システムを提案している. 一方で本研究では, 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

本章では提案する可視化システムの概要について述べる. 本システムは, 図1に示すように複数の可視化要素から構成されている. 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.

View 1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. 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. 全てユーザとコンテ

Visualization of Bias of Machine Learning for Content Recommendation

[†]Ami Tochigi and Takayuki Itoh, Ochanomizu University

[‡]Xiting Wang, Microsoft Research Asia

ンツそれぞれ 10 段階で配色を行う。

- (1) **Similarity**: 散布図上で選択したノードとのコサイン類似度に基づいて他ノードを配色。
- (2) **Difference**: 各ノードの学習データと推薦結果の属性分布を特徴ベクトルとし、それらのコサイン類似度に基づいて配色。
- (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[2]の手法を用いて各ノードに対する nDCG を算出。
- (4) **Coverage**: 推薦結果が網羅するジャンルの割合に基づいて配色。
- (5) **Binomial Diversity**: Vargas ら[3]が定義する網羅性と冗長性から多様性を評価する指標。

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。

View 2は、特定のクラスターの属性分布を示す棒グラフである。これによって、各クラスターに属するユーザやコンテンツの特徴や、偏った推薦の要因になりうるクラスターの発見が期待できる。さらに、**View 3**は特定のユーザノードあるいはコンテンツノードにおける属性分布を示す。ここでは、各ノードの学習データと推薦結果を比較表示することが可能である。**View 4**の帯グラフでは特定の属性についての分布を表示する。**View 5**と**View 6**は各ノードの詳細情報を表示するデータテーブルである。

4. 実行結果

4.1 データの前処理

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 によるモデル、BPR に GAN を適用したモデルの 2 種類を使用し、各ユーザに対して 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 **View 1** を **Difference** モードで配色し、各クラスターを観察すると、図 2 の散布図に示すように、彩度が高いノードが多いユーザクラスター 1 と 5 が確認できる。これは学習データと推薦結果の違いが大きいユーザが多いと考えられる。そこで、**View 2**を用いて各クラスターの属性分布を観察する。寒色の棒グラフはユーザ属性の統計、暖色の棒グラフはコンテンツ属性の統計を示す。

図 2 の棒グラフを見ると、それぞれ女性や子供が多いクラスターであることがわかる。ここで、女性や子供といった属性は使用データの中では少数である。よって、視聴結果と推薦結果の違いが大きくなる要因の 1 つにユーザ属性の偏りが考えられる。

次に上記クラスターから 1 人のユーザを選択し、**View 3**上で視聴結果と推薦結果を比較した。ここで、寒色の棒グラフは視聴結果、灰色の棒グラフは推薦結果を表す。図 3 を見ると、Children's など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。

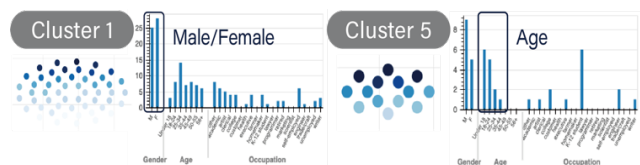


図 2 ユーザクラスターの属性分布

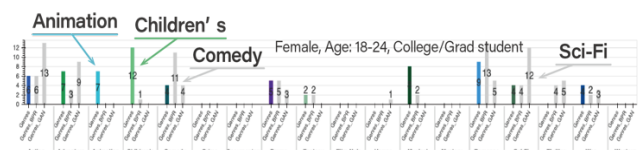


図 3 視聴結果と推薦結果の違いが大きいユーザの可視化結果

5. まとめ

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ることで、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。

参考文献

- [1] Y. Ahn and Y.-R. Lin. FairSight: Visual Analytics for Fairness in Decision Making. In IEEE Transactions on Visualization and Computer Graphics, Vol. 26, pp. 1086-1095, 2020.
- [2] 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.
- [3] 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.