5ZF-07

データバイアスを俯瞰するための階層型データ可視化 -可視化空調の温感の男女差への応用-

中井 祐希 † お茶の水女子大学 †
伊藤 貴之 ‡ お茶の水女子大学 ‡
高橋 秀和 § 富士通 §
中島 哲 ¶ 富士通 ¶
山本 哲 ‖ 富士通 ‖

1 はじめに

データのバイアスを発見するには，データが有する属性ごとの数値分布の違いを観察することが重要である. 特定の属性に起因する不利益を解消することは，公平性の高い社会の構築のためにきわめて重要である.

本報告では多数の人物を対象としたデータから，データの分布のバイアスを可視化する一手法を提案する. 人物群を属性で階層的に分割し，「平安京ビュー」[1] を用いて可視化する. ここで階層を表現する長方形領域を複数の帯グラフで塗りつぶし，その帯グラフによって所定の属性値に関する数値分布を表現することで，特定の属性に起因する数値分布の違いを可視化する. また，空調の温感に関する評価値の男女差を可視化した事例を報告し，本手法の有効性を議論する.

2 関連研究

2.1 データのバイアスの可視化

Cabrera らによる FairVis[2] は，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可視化解析システムである. 栃木ら [3] は，推薦システムにおける機械学習のバイアスを可視化している.

2.2 階層型データ可視化手法「平安京ビュー」

伊藤らによる「平安京ビュー」[1] は，階層型データの葉ノードを長方形のアイコンで，枝ノードを長方形の枠で表現し，階層構造を 2 次元の長方形群の入れ子構造で表現し，これらをできるだけ小さい画面空間に配置することで，階層型データ全体を一画面にする.

3 階層型データとしての偏りの可視化

提案手法では以下のデータが与えられることを前提とする. ここで A は人物集合によるデータ全体を表し，a_i は i 番目の人物を表し，n はデータ中の人数を表す.

$A = \{a_1, a_2, \ldots, a_n\}$

また，i 番目の人物に相当する a_i は以下の変数を有するものとする. ここで e_i は可視化の対象となる実数値，g_i は i 番目の人物の性別，r_{ij} は j 番目の実数型変数の属性値，c_{ik} は k 番目のカテゴリ型変数の属性値である.

$a_i = \{e_i, g_i, r_ij, \ldots, c_{ik}, \ldots\}$

提案手法では，ユーザが選んだ複数の属性値を用いて人物群を階層的に分類し，木構造を構成する. 提案手法ではこの木構造を「平安京ビュー」によって可視化する. 木構造の特定のノード配下に実数値 e_i の偏りが見られるようであれば，その偏りはユーザが選択した複数の属性値がもたらす交差バイアスに起因する偏りであることが示唆される.

「平安京ビュー」では葉ノードを正方形のアイコンで表現したのに対し，提案手法では葉ノード群に相当する人物群が有する e_i の分布を複数の帯グラフで表現する. 帯グラフの各領域の色算出は HSI 表色系を採用し，

Hierarchical data visualization of Gender Bias
-Application to Feeling of Temperature-
† Yuki Nakai, Ochanomizu University
‡ Takayuki Itoh, Ochanomizu University
§ Hidekazu Takahashi, Fujitsu
¶ Satoshi Nakashima, Fujitsu
‖ Tetsu Yamamoto, Fujitsu

色相 (H)　属性ごとに値を割り当てる.
彩度 (S)　平均値に近いほど低く，最大値/最小値に近いほど高く
明度 (I)　値が大きいほど高く.

という原則に沿って算出する.

4　空調温感データでの適用事例

本報告では空調の温感に関するオープンデータ [4] を適用した事例を示す. このデータから著者らは 32,373 人を対象として以下の属性値を抽出した.

TS　温感に対する評価値.0 がちょうどよい，正値が暑い，負値が寒い.
Sex　生物学的な意味での性別.
Age　年齢
Cloth　服装の厚さの実数値. 大きいほど厚い.
Metab　代謝量に関する実数値.
Season　春/夏/秋/冬のカテゴリ値.
Building　オフィス/教室/住居/高齢者施設/その他のカテゴリ値.
Strategy　エアコン/換気/混合のカテゴリ値.

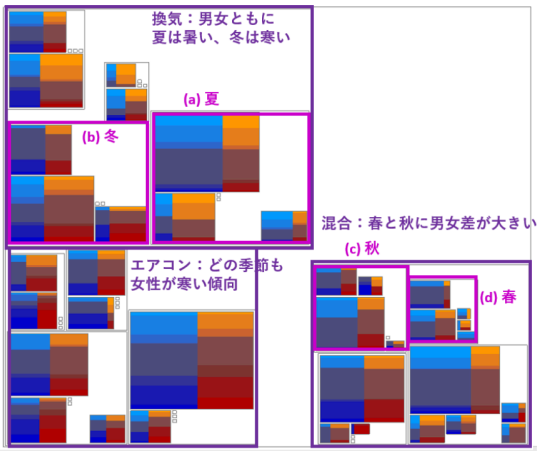

図 1　Strategy, Season, Building の順に人物を分類した例

図 1 は Strategy, Season, Building の順に属性値を参照して人物を分類した可視化結果である. 換気の枠の内側をみると，(a) 夏の枠では水色や橙色の領域が濃い青や濃い赤の領域よりも軒並み大きく，男女ともに「暑い」と判断している人が多いことがわかる. 逆に，(b) 冬の枠では水色や橙色の領域よりも濃い青や濃い赤の領域のほうが軒並み大きく，男女ともに「寒い」と判断する人が多いことがわかる. エアコンの枠の内側をみると，ほぼ全ての帯グラフにおいて濃い青より濃い赤の領域のほうが大きく，季節や場所を問わず女性のほうが「寒い」と判断する人が多いことがわかる. 混合の枠の内側をみると，(c) 秋の枠と (d) 春の枠において濃い青より濃い赤の領域のほうが大きい帯グラフのほうが多く，女性のほうが「寒い」と判断する人が多いことがわかる. 一方で夏と冬では男女間の判断の差が小さい傾向があることから，夏と冬ではエアコンと換気の混合が望ましいことが示唆される.

5　まとめ・今後の課題

本報告では多数の人物を対象としたデータ中に潜む交差バイアスを階層型データとして可視化する一手法を提案した. 空調の温感データを題材として可視化結果を示し，その有効性に就いて議論した.

今後の課題として，非常に多くの属性を有するデータにおいて，可視化する価値のある属性の組み合わせを自動選出する手法の開発や, 空調の温感以外の多様なデータへの適用が挙げられる.

参考文献

[1] T. Itoh, Y. Yamaguchi, Y. Ikehata, Y. Kajinaga, Hierarchical Data Visualization Using a Fast Rectangle-Packing Algorithm, IEEE Transactions on Visualization and Computer Graphics, Vol. 10, No. 3, pp. 302-313, 2004.
[2] Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, D. H. Chau, FairVis: Analytics for Discovering Intersectinal Bias in Machine Learning, IEEE Conference on Visual Analytics Science and Technology, 2019.
[3] A. Tochigi, T. Itoh, X. Wang, Visualization of Bias of Machine Learning for Content Recommendation, IEEE VIS, Posters, 2021.
[4] ASHRAE Global Thermal Comfort Database II, https://www.kaggle.com/datasets/claytonmiller/ashrae-global-thermal-comfort-database-ii