

ブログユーザの嗜好の可視化による集団的傾向の分析支援

小林 弘明^{1,a)} 古川 忠延^{2,b)} 三末 和男^{3,c)}

概要: ソーシャルメディア上では多種多様な属性のユーザが記事を書いており、その記事には各自の嗜好に関する内容が含まれている。本研究では、ブログのようなソーシャルメディアにおけるユーザの情報から実利性の高い知見を得ることを目的とし、ユーザ情報を詳細に分析するための可視化手法及びツールを開発した。開発した手法では、ユーザのデモグラフィックな属性と各趣味に対するスコアを平行座標軸上に表現することにより、ユーザ属性及び趣味スコアの関係性を知ることができる。実在ブログのユーザ情報をを用いたユースケースでは、本ツールによって発見できたブログユーザの集団的傾向を示した。

キーワード: 情報可視化, ソーシャルメディア分析, 多次元データ.

Analysis Support of Collective Trends by Visualization of Blog Users' Preferences

HIROAKI KOBAYASHI^{1,a)} TADANOBU FURUKAWA^{2,b)} KAZUO MISUE^{3,c)}

Abstract: Many different types of articles have been written on social media, containing content about the user's own preferences. We have developed a tool and a visualization technique for the purpose of obtaining highly pragmatic knowledge from user data in social media such as blogs. By representing both the user's demographic attributes and the score of each user preference on parallel axes, our technique allows us to find the relationship between the user attributes and the preference score. In a use case considering the user data of a real blog, we showed the collective trends regarding the blog users.

Keywords: Information Visualization, Social Media Analytics, Multi-Dimensional Data.

1. はじめに

ソーシャルメディア上では多種多様な属性のユーザが記事を書いており、その記事には各自の嗜好に関する内容が含まれている。また、ユーザは他のユーザが書いた記事を読むことができ、必要に応じてコメント等のフィードバ

クを与えることができる。このようなコミュニティが積み重なることによってソーシャルメディアは成り立っている。

ソーシャルメディアを解析して得たデータからは、“各属性のユーザが何に対して高い関心を持っているのか”という集団的傾向を知ることができる。ここから得られる知見は、マーケティング分野への幅広い応用が期待できる。例えば、ある商品売りたとき、その商品がどのような客層に好まれているかを知ることにより、顧客層を絞った効果的な広告を打つことができる。また逆に、客層毎の嗜好を把握することにより、顧客層が好みそうな商品やサービスを開発・提供することができる。このように、ソーシャルメディアから得られる集団的傾向は、利益を増やす上で重要な要素となり得るだろう。

本研究の目的は、ブログのようなソーシャルメディアに

¹ 筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻

Department of Computer Science, Graduate School of Systems and Information Engineering, University of Tsukuba

² 株式会社富士通研究所

Fujitsu Laboratories Ltd.

³ 筑波大学システム情報系

Faculty of Engineering, Information and Systems, University of Tsukuba

a) hiroaki@iplab.cs.tsukuba.ac.jp

b) furukawa.tada@jp.fujitsu.com

c) misue@cs.tsukuba.ac.jp

おけるユーザの情報から知見を得ることである。より実利性の高い知見を得るため、詳細な分析が可能な可視化ツールを開発する。

1.1 対象データ

本研究で扱うデータは、ブログユーザに関する情報を記したものである。ここで言うユーザの情報とは、ユーザのデモグラフィックな属性（ユーザ属性：表1を参照）及び各趣味に対するユーザのスコア（趣味スコア）である。趣味スコアは、“それぞれの趣味に関連する記事をどの程度書いているか”をユーザ毎・趣味毎に数値化したものである。各趣味について、関連する記事を多く書いているほど、その趣味のスコアが高くなるように集計した。趣味スコアの値域は $[0,1]$ の実数値であり、いずれの趣味もスコアの平均は0.5付近である。趣味の数については、Yahoo!カテゴリの趣味とスポーツ^{*1}を参考に、25種類に分類した。

表1 対象データにおけるユーザ属性のカテゴリ一覧

Table 1 Category List of user attributes in the target data

属性名	カテゴリ名 (ユニーク数)
ユーザ ID	数値 (5000 値)
性別	男性, 女性, 不明 (3 値)
結婚歴	未婚, 既婚, 不明 (3 値)
職業	中高生, 大学専門学校生, 社会人主婦, シニア, その他 (6 値)
居住地	東京, 大阪, 愛知, その他 (4 値)

データは tsv 形式で保存されており、各行（レコード）が各ブログユーザ、各列が各ユーザ属性及び趣味スコアに対応している。今回は 5000 レコードのデータ、つまり 5000 人分のブログユーザのデータを扱うことを考える。このデータは 5 種類のユーザ属性と 25 種類の趣味から成るデータであり、これは 30 次元の多次元データとして捉えることができる。本研究では、多次元データの可視化手法を用いて対象データを可視化することにより、データの分析を行うこととする。

1.2 分析要求

対象データに対する主な分析要求は、任意に抽出したユーザの部分集合における特徴を把握することである。主な集団の特徴として、我々は以下の 3 点に着目した。

- (1) 集団がどのような趣味に対して高い関心を持っているか。
- (2) ある趣味に高い関心を持っている集団は、他にどのような事に関心を持っているのか。
- (3) ある趣味に高い関心を持っている集団は、どのようなユーザ属性から成るのか。

^{*1} <http://dir.yahoo.co.jp/Recreation/>

これらの特徴を効果的に把握できる表現手法及びツールを開発する。

2. 関連研究

本研究の対象データは多次元データとして扱うことができる。そこで、多次元データの可視化手法を本研究の関連研究として紹介する。

2.1 Cartesian Displays

Cartesian Displays は直交座標系を利用した可視化手法である。Cartesian Displays を用いて多次元データを可視化するため、複数の可視化結果のパネルを行列状に並べる手法（Panel Matrix）が研究されてきた。

代表的な Panel Matrix の可視化手法としては、散布図行列 [1] が挙げられる。散布図行列は、全ての組み合わせ可能な次元対の散布図を行列状に並べて表示する。例えば N 次元データの場合、 $N(N-1)$ 通りの散布図を表示することにより、データ全体を俯瞰できる。

SCATTERDICE[2] は散布図行列を拡張した可視化手法で、多次元データにおける各次元を切り替える作業を、サイコロを転がすようにインタラクティブに操作することで実現する。通常は 2 次元の散布図を表示し、次元を切り替える際には、奥行きにあるもう 1 次元との散布図が 3 次元のアニメーションによって変遷する。これにより、一般的な 2 次元の散布図が持つシンプルさを活かしつつ、次元の切り替えを直感的に行うことができる。

Mosaic Plot[3] は Cartesian Displays の一種であり、空間充填型のカテゴリデータ可視化手法である。この手法では、描画領域を矩形に分割することにより、データ分布を各矩形の面積で表現する。この矩形の分割数を増やすことにより、多次元のカテゴリデータを表現することができる。しかし 3 つ以上の次元を表現すると、矩形の数が多くなってしまい、読解が難しくなる。

色付き Mosaic Matrix[4] は、Mosaic Plot を拡張して行列状に配置した、Panel Matrix の多次元データ可視化手法である。色付き Mosaic Matrix はデータの特徴を色を用いて表現することにより、限られた描画領域内でもデータの概観を読み取ることが可能な表現手法である。

以上の Cartesian Displays を用いた Panel Matrix の可視化手法は、個々の表現自体がシンプルであるため、読解が容易であるという利点がある。一方で、これらの手法の欠点としては、複数の次元が関連している傾向や特徴の発見が困難である点が挙げられる。一般的に、1 つの直交座標系において次元数の多い高次元データを表現するのは困難であり、たとえ表現したとしても読解が困難である。また、各 2 次元対の可視化結果を Panel Matrix として並べたとしても、直接的には 2 次元対の特徴しか見ることができないため、複数の次元にまたがる関係性を読み取ること

は難しい。

2.2 Non-Cartesian Displays

Non-Cartesian Displays は直交しない座標系を利用する可視化手法である。代表的な Non-Cartesian Displays の多次元データ可視化手法としては、座標軸を平行に並べる Parallel Coordinates Plot (PCP) [5] が挙げられる。PCP は多次元の量的データを効果的に可視化できる手法である。PCP では、まず各次元に対して座標軸を用意し、それらを並列に並べる。そして全てのレコードに対して、各座標軸における同一レコードの点を結んでいく。これによって、1本の線で1つのレコードを表現する。PCP は、概観の把握や少数のレコードに焦点を当てた分析に対して効果的な手法である。一方で、PCP は線の集合による概形によってデータを表現するため、詳細な読み取りは困難である。また、各座標軸では点で値を表現するため、カテゴリデータの分布の読み取りも難しい。

Angular Histograms[6] は、多次元データの各次元についてヒストグラムを作成し、PCP と同様に平行に並べて表示する手法である。さらに元となる PCP の線の傾きを元に、ヒストグラムの各棒の傾きを設定することで、隣り合う次元同士の関係性を表現する。Angular Histograms では、各値の頻度をより詳細に読み取ることが可能である。しかし、表現が隣の軸に依存してしまうため、離れた軸同士の比較が困難である。またデータの次元数が増えると読解が困難になる。

Parallel Sets[7] は、PCP と Mosaic Plot を組み合わせた多次元カテゴリデータの可視化手法である。PCP における座標軸上の各点（またはカテゴリ）を大きさを持った矩形として表現し、さらに矩形の幅を持った線として座標同士を繋ぐことによって多次元データを表現する。またカテゴリに応じて矩形の色を設定することで、複数次元間の関係性を表現できる。しかし Parallel Sets はカテゴリデータに特化した可視化手法であるため、量的データの可視化及び分析は困難である。

3. 視覚的表現

3.1 表現手法の設計

対象データに対する分析要求（1.2 節を参照）を満たすため、表現手法の要件を以下の通りに設定した。

- (1) **データ分布の読み取りが容易である。** 集団がどのような趣味に対して高い関心を持っているかを知るためには、各趣味におけるブログユーザのスコア分布を読み取れる必要がある。
- (2) **複数の次元間における関係性の読み取りが容易である。** 趣味間の関係性を知るためには、趣味同士でスコア分布の比較を行う必要がある。また、趣味間の相関関係についても把握できることが望ましい。

- (3) **複数の尺度水準が混在したデータに対しても分析可能である。** 対象データにおけるユーザ属性はカテゴリデータであり、趣味スコアは量的データである。そのため、ユーザ属性と趣味スコアの関係性を知るためには、異なる尺度水準の情報を併用した分析を行う必要がある。

以上を考慮して、我々は平行座標軸上にデータを表現する手法を設計する。

対象データは 5 種類のユーザ属性及び 25 種類の趣味スコアで構成されており、これは 30 個の次元を持つ多次元データと捉えることができる。そこで、1つの次元に対して1本の座標軸を割り当てた上で、全ての座標軸を平行に並べて表示する。つまり、30本の平行座標軸によって対象データの全体を表現する。

また本表現手法では、ユーザ属性と趣味スコアの表現においてそれぞれ異なる手法を用いる。これにより、複数の次元間にまたがる比較や関係性の読み取りを可能にする。

3.2 ユーザ属性の表現

ユーザ属性は、図 1 のように、100 % 積み上げ棒グラフを用いて表現する。1つのユーザ属性が1本の棒グラフに対応しており、棒内の各矩形はそれぞれ1つのカテゴリに対応している。矩形の高さはカテゴリの比率に応じて決定している。これにより、属性内のカテゴリの構成を読み取ることができる。

例えば、図 1 の左から 2 本目の棒グラフには、ユーザの性別が割り当てられている。男性（1435 人）を赤、女性（3270 人）を黄色、不明（295 人）を緑の矩形によってそれぞれ表現している。

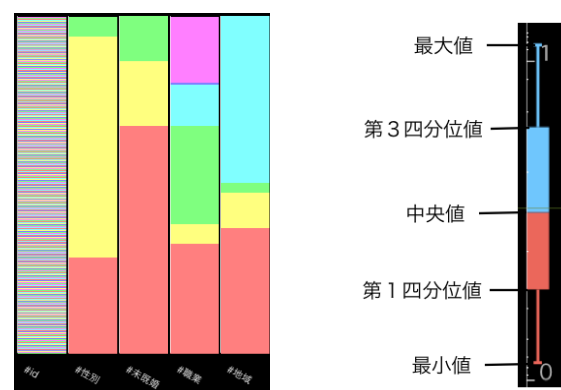


図 1 ユーザ属性の表現に用いる 100%積み上げ棒グラフの例。 図 2 趣味スコアの表現に用いる Boxplot の例。

Fig. 1 Example of 100%

stacked bar chart for the representation of user attributes.

Fig. 2 Example of Boxplot for the representation of the preference score.

3.3 趣味スコアの表現

趣味スコアは、図2のように、Boxplot [8] を用いて表現する。Boxplot では、矩形および上下に伸びる直線を用いて5種類の要約統計量を表現する。この5種類の要約統計量はそれぞれ、最小値、第1四分位値、中央値、第3四分位値、最大値である。Boxplot はデータの要約統計量のみを表現するため、視覚的混雑度を抑えた上で、データ分布の読み取りに必要な情報を示す事が可能である。

Boxplot を並列に並べることの利点として、PCP による表現と比べて、複数の次元間にまたがるスコア分布の比較がより容易であるという点が挙げられる。比較の目安となる5種類の統計量について、その位置を比べることにより、簡単かつ素早く複数の趣味スコアの分布を比較することができる。

3.4 Parallel Coordinates Plot によるユーザ表現

Boxplot は表現可能な情報量が少ないため、そのままでは趣味間の相関関係などを表現できない。そこで本ツールでは、図3のように、PCP を用いて1人のユーザを1本の線で表現可能にした。分析の際に PCP を併用することによって、Boxplot では表現できない細かなスコア分布を確認することが可能になる。また他にも、趣味間の相関や特徴的なスコア分布を知ることができる。

さらに PCP によって表現する情報量を増やすため、レコード毎に線の色を設定する。全趣味に対する平均スコアをレコード毎に計算し、その値に応じて線の色を決定する。あるレコード i の趣味スコアの平均スコアを $\mu_i (0 \leq \mu_i \leq 1)$ とおく。多くの趣味に対して多くの記事を書いているブログユーザほど、 μ_i が1に近くなる。まず色相について、 $\mu_i < 0.5$ であれば赤、 $\mu_i \geq 0.5$ であれば青とする。次に彩度 S_i について、 $S_i = \sin(\pi|\mu_i - 0.5|)$ とする。以上により、 μ_i が0に近いユーザは赤、0.5に近いユーザは白、1に近いユーザは青の線によってそれぞれ描画される。

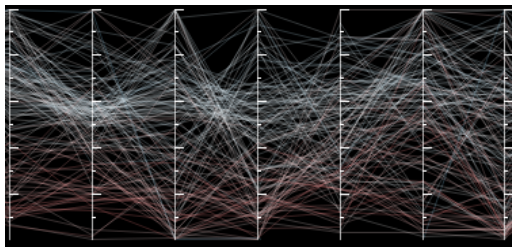


図3 PCP を用いたユーザ表現の例。

Fig. 3 Example of the user representation by PCP.

4. ツールの開発

3節で述べた表現手法を用いて、対象データを可視化して分析するツールを開発した。ツールのスクリーンショットを図4に示す。本ツールでは、30本の平行座標軸を用い

てデータを可視化している。ツール画面の左側5本の座標軸は、100%積み上げ棒グラフによりユーザ属性を表現している。その他の25本の座標軸は、各趣味スコアの分布を表現している。なお、Boxplot と PCP に関しては、表示/非表示を切り替えることが可能である。図4では、Boxplot と PCP を同時に表示している。

各座標軸の下には、割り当てられているユーザ属性または趣味のラベルが記されている。また、各値のラベルはマウスオーバーによって表示される。これにより視覚的混雑度を減らし、全体を俯瞰しやすくしている。

4.1 実装言語及びデータ形式

本ツールの実装言語には Java(Java™ Platform Standard Edition 6.0) 及び Processing*2を使用した。processing.core.PApplet を継承することで Java のプログラムに Processing のスケッチを埋め込んでおり、描画部分の実装に Processing を使用している。

本ツールは tsv 形式のデータを外部から読み込む。読み込むデータファイルは以下の2つである。

- **ブログのユーザ情報が記述されたファイル (データ本体)**。1行が1人のブログユーザ、1列が1つのユーザ属性または趣味スコアに対応する。全ての行について、各列に必ず1個の値を持つものとする。なお、ユーザ属性の次元は String 型、趣味スコアは Float 型で読み込む。
- **ユーザ属性及び趣味のラベルが記述された1行のファイル**。これはデータ本体において、各列が何の情報であるかを記したファイルである。ラベルの間はタブで区切り、並びは『ユーザ属性の各ラベル』→『趣味の各ラベル』の順とする。先頭に半角の#が付いているラベルをユーザ属性として処理する。#がなくなった時点で、それ以降のラベルを趣味として処理する。

読み込むファイルを分割することにより、データ本体のみに変更が生じた場合にも柔軟に対応できる。例えば表示するユーザ数を増やしたい場合は、データ本体に新たに行を追加するだけで、変更を反映させることが可能である。

4.2 レコードの選択機能

本ツールでは、レコードを選択することにより、レコードの部分集合を作成することができる。各座標軸における値をクリックまたはドラッグすると、その値を含むレコードが選択される。レコードの選択においては、同一座標軸内の複数の値や値域、異なる座標軸の値を条件にすることが可能である。例えば、『性別が女性で、職業が中高生で、趣味Aに対して0.7以上のスコアを持つレコード』などという条件を設定できる。なお、軸内の複数箇所をドラッグ

*2 <http://processing.org/>

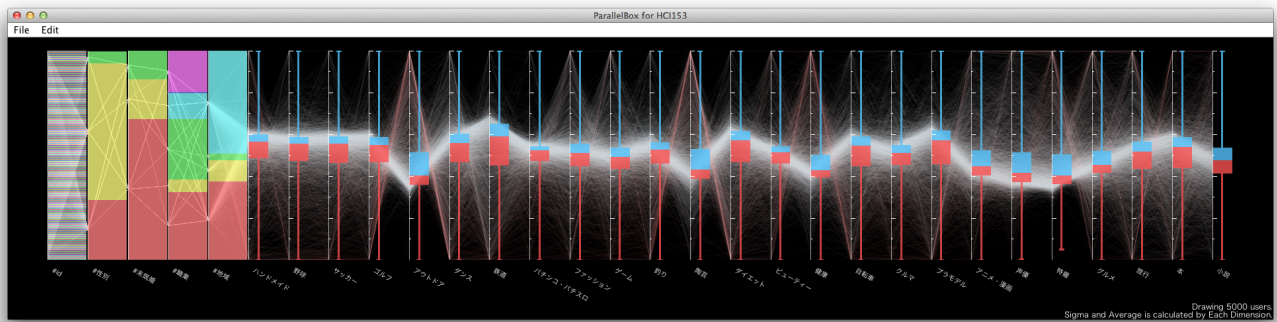


図 4 開発したツールのスクリーンショット。
Fig. 4 Screenshot of the developed tool.

することにより複数の条件を設定した場合は、それぞれの条件に合致するレコード集合の和集合となる。一方、複数の座標軸に条件が設定されている場合は、それぞれの集合の積集合となる。

任意の部分集合を作成すると、母集団との比較が可能な表現へと変化する。図 5 は、レコード選択時におけるツール画面の拡大図であり、性別が女性であるプログユーザのレコードのみを選択した状態を示している。

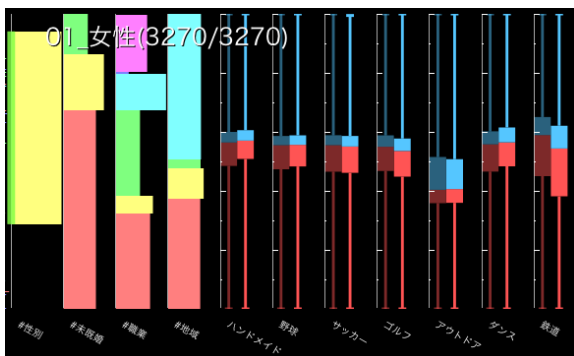


図 5 レコード選択時におけるツール画面の拡大図。
Fig. 5 Enlarged view of the tool in selecting records.

レコード選択時の 100%積み上げ棒グラフは、図 5 の左側のように、被選択レコードの含有率に応じて横幅が変化する。これにより、各カテゴリを値に持つレコードのうち、どの程度の割合のレコードを選択しているのかを知ることができる。例えば図 5 の左から 3 本目の軸は、職業のユーザ属性を表現している。この軸の緑の矩形は社会人に対応しており、この矩形の幅が最大時の約 1/2 になっていることから、このデータに含まれている社会人の約半数は女性であることがわかる。また水色の矩形は主婦に対応しており、この矩形の幅から、主婦は全員女性であることが読み取れる。

レコード選択時の趣味スコアの表現として、図 5 の右側のように、母集団に対する Boxplot と部分集合に対する Boxplot を並べて表示する。2 つの Boxplot は明度の違い

によって区別が可能であり、前者の方が暗い色となっている。2 つの Boxplot を表示することにより、レコード選択前後での各趣味のスコアを比較することができる。例えば図 5 は性別が女性のレコードを選択した状態の図であり、最も右の軸は鉄道の趣味スコアを表現している。鉄道の軸にある 2 つの Boxplot を比較すると、明るい方（女性ユーザの集合）は暗い方（母集団）と比べて、第 1 四分位値及び中央値、第 3 四分位値の値が低くなっていることが確認できる。ここから、女性は鉄道に関する関心が低い傾向にある、ということがわかる。

本ツールには、被選択レコードを新しい母集団として、その可視化結果を新規ウィンドウ上に表示する機能を備えている。これにより、ドリルダウン式の詳細な分析を可能にしている。

4.3 趣味軸のソート機能

座標軸の下に表示されている趣味のラベルをクリックすることで、その趣味を選択状態にすることができる。この時、他の趣味と選択状態の趣味との相関係数を計算し、ラベル上にそれぞれの相関係数を表示する。また、選択状態の趣味をさらにクリックすることで、図 6 のように、相関係数の高い順に趣味の軸を並び替えることが可能である。これにより、ある趣味に着目した時、その趣味に対して強い正の相関を示している趣味を簡単に見つけることができる。



図 6 相関係数を用いた趣味ソートの例。
Fig. 6 Example of sorting interests by correlation coefficient.

各ラベルの色は、計算された相関係数に対応付けされている。正の相関が強いほど青、相関が弱いほど白、負の相関が強いほど赤色のラベルとなる。なお、選択状態の趣味は緑である。

5. ユースケース

開発したツール及び 1.1 節で述べたデータを用いたユースケースを示す。なお、図 4 は対象データの可視化結果である。

5.1 趣味スコアからの分析結果

まずは PCP を表示して、データ全体を俯瞰した。すると複数の趣味において、図 7 のような“赤色の山”を確認することができた。具体的には、アウトドア、陶芸、健康、アニメ漫画、声優、特撮に対応する軸において、この“赤色の山”を確認した。PCP における赤線は、“趣味スコアの平均値が低いレコード”を意味している。ここから、“一部の趣味に関する記事ばかり書いていて、他の趣味に関する記事をほとんど書いていないために、趣味の平均スコアが低くなっているブログユーザ”が多く存在していることがわかった。

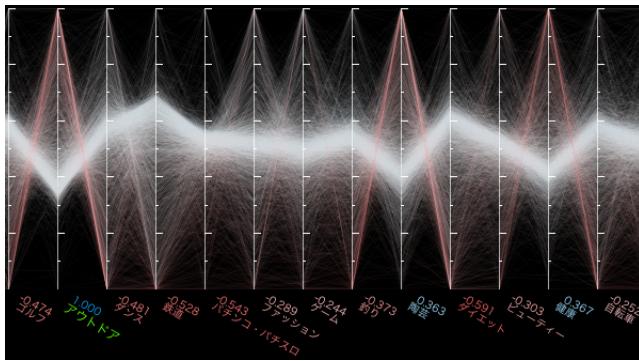


図 7 PCP のみの表示結果。
Fig. 7 Visualization result by PCP only.

最も顕著な“赤色の山”が現れたアウトドアに着目し、アウトドアに対する相関係数で趣味の軸をソートした。こゝで表示を Boxplot に切り替え、さらにアウトドアのスコアが第 3 四分位値以上のレコードのみを選択した。すると、先述の“赤色の山”を確認した趣味については、中央値、第 3 四分位値が高いスコアであることが確認できた。ここから、これらの趣味は関連性が高く、またこれらを総じて好むブログユーザが多く存在する、ということがわかった。

次に、アニメ・漫画に関心が高い人に着目して分析を行った。アニメ・漫画に対して各趣味の相関係数を計算したところ、特撮や声優、ゲームが強い正の相関を示していることがわかった。アニメ・漫画におけるスコアが第 3 四分位値以上であるレコードのみを選択したところ、上記の趣味に加えて、小説のスコアが比較的高いという傾向が見られた。この時 100%積み上げ棒グラフからユーザ属性を読み取ったところ、男女比は母集団とほぼ変わらず、中高生が比較的多い、といった傾向が確認できた。

また、サッカーとゲームに着目して、両方に関心の高い

ブログユーザが他にどのような事に興味があるかを分析した。まずサッカーとゲームが共に第 3 四分位値以上のスコアであるレコードのみを選択し、画面を再描画した。図 8 は、再描画した結果の一部である。図 8 より、サッカーとゲームが好きなブログユーザは、アニメ・漫画や声優、野球、ダンスなどにおいても高いスコアである場合が多かった。またユーザ属性を確認したところ、中高生が他の職業と比べて圧倒的に多いことがわかった。

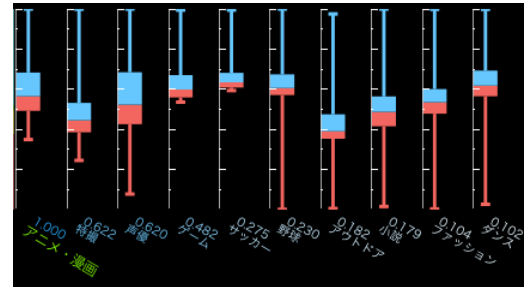


図 8 サッカーとゲームに関心の高いブログユーザのレコードのみで再描画した結果。

Fig. 8 A result of re-drawing only records of blog users who are strongly interested in games and football.

5.2 ユーザ属性からの分析結果

ユーザ属性から分析を進めることにより、顧客層を絞った分析を行うことができる。本節では、ユーザ属性からの分析結果について 3 種類の結果を紹介する。

5.2.1 女子中高生 (1111 人) に着目した分析

まず、性別が女性かつ職業が中高生であるブログユーザのレコードを選択した。Boxplot を比較したところ、女子中高生は母集団と比べて、アニメ・漫画、声優、ダンスなどにおいて比較的高いスコアであることが確認できた。一方で、アウトドアや健康、陶芸などは比較的低いスコアであった。ここで、女子中高生のレコードのみで再描画を行い、さらにダンスのスコアが第 3 四分位値以上であるレコードを選択した。図 9 はその時の画面の一部である。図 9 より、ダンスに関心の高い女子中高生は、サッカーや野球などの運動系のスコアも高いため、体を動かすことが好きであると推測できる。また、ファッションや声優、アニメ・漫画などにも高い関心を持っていることが確認できた。

5.2.2 主婦 (622 人) に着目した分析

図 10 は職業が主婦であるレコードを選択した結果である。まず特筆すべき点として、PCP において多くの赤線が確認でき、また多くの“赤色の山”が確認できた事が挙げられる。これは、多くの主婦にとってブログの用途が限定的であり、記事の内容が一部の趣味に関するものばかりであることを示唆している。次に Boxplot によって各趣味スコアを母集団と比較すると、ハンドメイドやアウトドア、陶芸、健康、グルメにおけるスコアが比較的高いことがわ

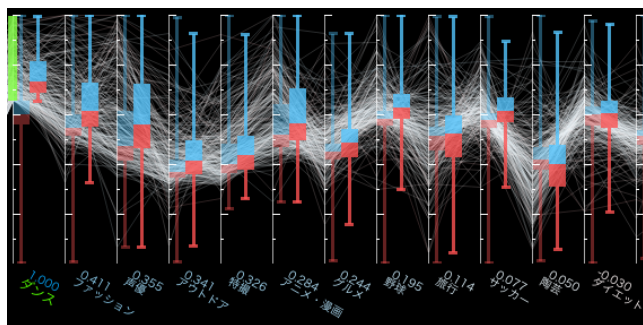


図 9 女子中高生の中で、ダンスのスコアが上位であるブログユーザのレコードを選択した結果。

Fig. 9 A result of selecting records in the women's teenagers whose score of dance are higher.

かった。一方で、運動や遊びに関する趣味については比較的低いスコアであった。

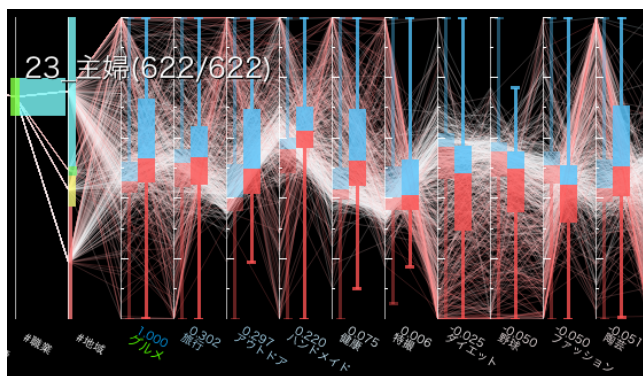


図 10 主婦のみを選択した結果の一部。

Fig. 10 Part of the result of selecting the only housewife.

5.2.3 社会人 (1449 人) に着目した分析

まず職業が社会人であるレコードのみを選択して再描画を行った。再描画後、既婚のみを選択した結果が図 11 である。社会人全体と比べて、アウトドア、特撮、健康、陶芸などのスコアが比較的高いことが確認できた。ここから、既婚の社会人は休日の息抜きや子供と遊んだことをブログの記事にしているのではないかと推測できる。

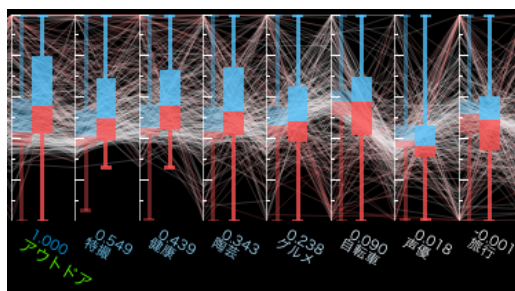


図 11 社会人のうち既婚のみを選択した結果。

Fig. 11 A result of selecting married members of society.

5.3 議論

本ツールはマーケティングへの活用、例えば商品の売上数を増やすことを主眼に置いて開発したものである。商品の売上数を増やせるか否かは、効果的な宣伝を行えるか否かが重要な鍵となる。その商品に対して、関連性の高いトピックや関心の高い客層を探し出すことにより、顧客層を絞った宣伝が可能になる。このような宣伝を可能にするための知見は、本ツールを用いて趣味スコアから分析を行うことによって得ることが可能である。例えばサッカーゲームに関する情報を集めたい場合、実際にユースケースで示したとおり、サッカーやゲームに関心のあるブログユーザー層や、それらと関連性の高いトピックを探すことができる。

商品やサービスを提供するためには、まずその商品がターゲットにしている顧客層の嗜好を把握することも重要である。性別や職業を始めとしたデモグラフィックな情報は、それぞれの層の嗜好へも影響を及ぼすと考えられる。本ツールを用いることにより、ユーザ属性からの分析によって知見を得ることが可能である。また本ユースケースでは、ブログのユーザー層を絞った分析についても言及した。

以上により、本ツールを用いてブログユーザーに関する情報を分析することによって、マーケティング等に活用可能な知見を得られることを示した。本ツールにより得られた知見は、以下の様な活用が期待できる。

- ある趣味を持った顧客層が他にどんなことに趣味を持っているかが分かり、新たな商品を推薦（宣伝）できるようになる。
- 特定の職業や性別における趣味の傾向を知ることにより、顧客層に合った商品・サービスを開発するヒントが獲得できるようになる。

本ツールに用いた表現手法について、趣味スコアの表現に PCP と Boxplot を併用することによって、より詳細な分析を進めることが可能となった。Boxplot 単体では、スコア分布の比較が容易である一方で、表現可能な情報量が少ないという欠点がある。PCP 単体では、各レコードの趣味スコアや趣味間の関係性を読み取ることは可能であるが、異なる趣味間でのスコア分布の比較は困難である。2つの手法を併用することにより、それぞれの手法が持つ欠点を補うことが可能となった。

一方で、本ツールには比較可能な集合が限定的であるという限界がある。より幅広い分析を可能にするためには、より多種類の比較ができるような表現及びツールの機能が必要であると考えられる。また、本ツールにおけるデータの比較の際には、人間が知覚した情報を判断しなければならない。より複雑なデータを扱うためには、より比較が容易な表現を用いる必要があるだろう。

6. まとめと今後の展望

本研究では、ブログのようなソーシャルメディアにおけ

るユーザの情報から知見を得るための表現手法及び分析ツールを開発した。開発した表現手法には、ユーザ属性と趣味スコアに対して異なる表現手法を用いた。両表現手法を平行座標軸上に表現することにより、ユーザ属性及び趣味スコアの関係性を表現することが可能になった。また、本表現手法を用いて分析ツールを開発した。より実理性の高い知見を得るため、複数の観点から分析を行うことが可能な機能を実装した。また、開発した分析ツールを用いたユースケースを示した。実際のデータに対する分析を行うことで、確認できた知見を紹介すると同時に、本ツールの有用性を示した。

今後の課題としては、ツールに利用する表現手法やインタフェースの改良、及び実用可能なシステム化が挙げられる。将来的には Web ブラウザ上でのインタラクティブな分析操作が可能となることが望ましいと考える。また、ブログ以外のソーシャルメディアにおけるユーザ情報を収集し、そのデータを分析することにより、今回とは異なる知見を得ることができだろう。

参考文献

- [1] D. B. Carr, R. J. Littlefield, W. L. Nicholson and J. S. Littlefield. Scatterplot Matrix Techniques for Large N. In *Journal of the American Statistical Association*, Vol. 82, No. 398, pp. 424–436, 1987.
- [2] N. Elmqvist, P. Dragicevic and J.-D. Fekete. Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 14, No. 6, pp. 1141–1148, 2008.
- [3] M. Friendly. Mosaic Displays for Multi-Way Contingency Tables. In *Journal of the American Statistical Association*, Vol. 89, No. 425, pp. 190–200, 1994.
- [4] 小林弘明, 三末和男, 田中二郎. 色による特徴表現を用いた高次元データの可視化情報処理学会研究報告 (152 回ヒューマンコンピュータインタラクション研究会), Vol. 2013-HCI-152, No. 23, 2013.
- [5] A. Inselberg and B. Dimsdale. The plane with parallel coordinates. *The Visual Computer*, Vol. 1, No. 4, pp. 69–91, 1985.
- [6] Z. Geng, Z. Peng, R. S. Laramée, R. Walker, and J. C. Roberts. Angular Histograms: Frequency- Based Visualizations for Large, High Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 17, No. 12, pp. 2572–2580, 2011.
- [7] F. Bendix, R. Kosara and H. Hauser. Parallel Sets: Visual Analysis of Categorical Data. *IEEE Symposium on Information Visualization (InfoVis2005)*, pp. 133–140, 2005.
- [8] Y. Benjamini. Opening the Box of a Boxplot. In *Journal of The American Statistician*, Vol. 42, No. 4, pp. 257–262, 1988.