人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第18回)
SIG-AM-18-06

# 着目点の明示によるデータ分析支援

# Data Analysis Support by Displaying Points to Notice

中川拓郎　砂山渡　畑中裕司　小郷原一智
Takuro Nakagawa　Wataru Sunayama　Yuji Hatanaka　Kazunori Ogohara

滋賀県立大学 工学部
School of Engineering, The University of Shiga Prefecture

**Abstract:**
現在，爆発的に増大しているデータを分析し価値のある知識を発掘するニーズが高まっている．しかし，膨大で捉え所のないデータの分析に際して，目をつけるべき着目点を見出すことは簡単ではない．そこで本研究では，データの中から分析の手がかりとなる着目点をハイライトにより明示をすることで，データ分析の支援を行う枠組みを提案する．データ分析において平均等の基準値からのズレが大きいデータを着目点として明示するとともに，与えた着目点を手がかりとして，データへの着目と絞り込みを繰り返し行える機能を実装した．着目点の明示機能を利用することで，データ分析の支援が可能であるかを評価実験を行い，提案システムの推奨する着目点の明示機能を利用することで，分析者のデータ解釈の数が増加する傾向がみられた．

## 1　緒論

現在，爆発的に増大しているデータを分析し価値のあるデータを発掘するニーズが高まっている [1]．しかし，膨大で捉え所のないデータから，目をつけるべき着目点を発見することは簡単ではない現状がある．

この問題点に関して，多量のデータの中から着目点を探す事が困難であるデータマイニングに慣れていない人に対して基本的な着目点を見出す支援を，データ分析に慣れている人には，データの見落としを最小限にする幅広い着目点を見出す支援がそれぞれ必要である．

そこで本研究では，テキストマイニングのための統合環境 TETDM （Total Environment for Text Data Mining）[2] をベースとして，データから平均や頻度の基準値からのズレが大きいデータを着目点として明示するとともに，与えた着目点を手がかりとして，データへの着目と絞り込みを繰り返し行える機能を実装することにより，初心者でも分析を行いやすいデータ分析環境の構築を行う．

## 2　関連研究

### 2.1　データ分析の着目点に関する研究

レビュアがレビューの評価値を決定した根拠となる商品の機能や特徴の提示を行う研究がある [3]．この研究では，レビュー文章中の商品に関する文章を抽出し，使われている形容詞について着目することで，日本語評価極性辞書からレビュー評価の根拠を示している．本研究では，レビュー全体から評価値や単語でハイライトと絞り込みを行うことでレビュー評価の根拠を解釈できる着目点の明示を行う点で異なる．

アンケートデータの解析時に他の人の回答との関連度が低い少数回答を明示する研究研究がある [4]．この研究では少数回答，少数意見に重点をおいて着目点を示しているが本研究では，さまざまな目的に対応して幅広く着目点になり得るデータを明示を行う点で異なる．

### 2.2　データの可視化に関する研究

テキストマイニングによる授業評価アンケートの分析時に共起ネットワークを用いた自由記述の可視化を行う研究がある [5]．この研究では，アンケートの自由記述から取り出した頻出語の 30 単語を用いて共起ネットワークによる可視化を行っている．本研究ではデータ中から分析者が示した着目点を直接ハイライトする可視化を行う点で異なる．

## 3　データ分析における着目点の明示機能

### 3.1　データ分析の流れ

図 1 にデータ分析の流れを示す．まず，分析を行うデータを入力する．入力した膨大な量のデータは一度

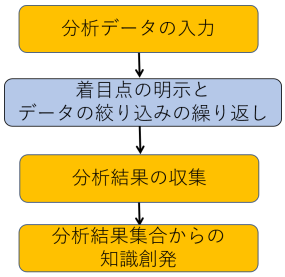


図 1: データ分析の流れ

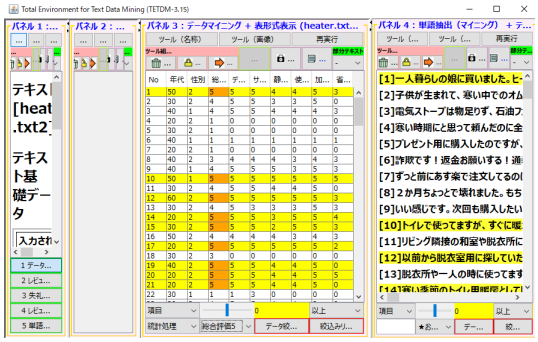

図 2: TETDM の画面の例

にすべてのデータを理解することができない．そのため，分析者が条件を選択しデータの明示，絞り込みを繰り返し行い，データを少なくすることで分析結果を見出す．分析結果集合からデータを整理，統合することで知識創発を行う．

データ分析を行うためには、データの絞り込みと繰り返しによってたくさんの分析結果を集めることが重要である．

## 3.2 データ分析環境：TETDM

データ分析環境の基盤として，テキストデータマイニングのための統合環境 TETDM[1]を利用した．

TETDM は，ツールを独自に開発，追加できる特徴を持つデータ分析の統合環境である．テキスト分析のための多様なツールを有し，テキストデータと数値データの両方を扱える．

TETDM の画面例を図 2 に示す．TETDM の 1 画面は，複数のパネルから構成されており，各パネルにデータ処理のためのツールと，データ可視化のためのツールをペアでセットして利用する．TETDM には，40 以上の処理ツールと可視化ツールが実装されており，それらを柔軟に組み合わせて利用することができる．

[1]http://tetdm.jp からダウンロード可能

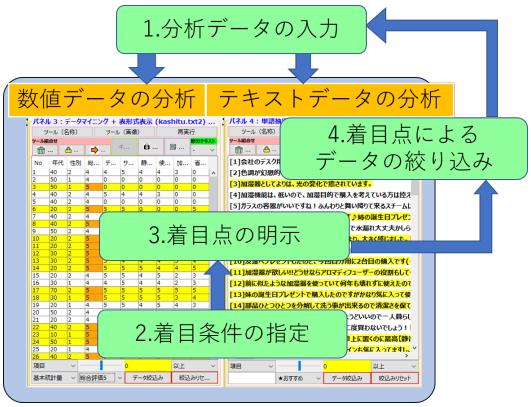


図 3: 着目点の明示とデータの絞り込みの流れ

## 3.3 データ分析における着目点の明示機能の枠組み

松本ら [6] により，TETDM を用いたデータ分析時に，数値分析ツールとテキスト分析ツールを連携して利用する環境が提案されている．しかし，データ分析時にどこに着目点してデータを分析するかという点について，分析者が自発的に着目点を発見する必要があった．

本研究では，何らかの条件に合致する一部のデータに対して着目点を明示し，データの絞り込みを繰り返し行う機能をデータ分析ツールとテキスト分析ツールに追加することで，分析者がデータ中の傾向や特徴を見出す支援を行う．

図 3 に，本研究で TETDM に追加した着目点の明示とデータの絞込み機能の流れを示す．すなわち，入力された分析データに対して，着目条件を設定し，その条件にマッチするデータを明示する．また，明示したデータをより深く分析するために，データの絞り込みを行って分析を繰り返す．この着目条件を指定する際に，条件の設定を手動で行いづらいという問題点を解決するために，おすすめの着目条件を提示する．以下の節で，この各ステップについての詳細を述べる．

## 3.4 分析データの入力

入力データは，数値データとテキストデータの両方，またはいずれか一方のみのデータセットを入力とする．各データは，属性と属性値のペアで構成されている．

TETDM においてテキストは以下のように分割して処理される．すなわち，テキストデータ全体を「文章」，テキストデータに挿入される「スナリバラフト」というタグで区切られた部分テキストを「セグメント」，句点で区切られた文を「文」として処理する．入力データの例を表 1 に示す．この例のようなデータにおいて

表 1: 本分析環境の入力データの例

| ID | 年代 | 性別 | 総合得点 | テキスト |
|---|---|---|---|---|
| 1 | 50 | 2 | 5 | 一人暮らしの娘に買いました。ヒーターは小さくて軽いので片手で運べるしちゃんと暖まれるので良いです！w だそうです。 |
| 2 | 40 | 2 | 3 | 2か月ちょっとで壊れました。もちろん保証で新しい物と交換できました。 |
| 3 | 40 | 1 | 4 | いい感じです。次回も購入したいと思います。 |
| 4 | 50 | 1 | 5 | トイレで使ってますが、すぐに暖かです。。 |
| 5 | 60 | 2 | 5 | 以前から脱衣室用に探していたところ、娘の家で使っていて快適だとのお墨付きをもらい購入しました。大満足です。 |
| 6 | 30 | 2 | 4 | 脱衣所や一人の時に使ってます。センサーがついてるのでつけっぱなしにならないので、電気代の節約になってると思います。デザインもシンプルで気に入っています。値段もお手頃でした。 |
| 7 | 30 | 2 | 5 | 夜中の授乳用に購入。すぐに温風が出るので短時間であたたかくなるので良かったです。 |

は，一人分のデータごとに「スナリバラフト」を挿入して，1 つのセグメントとして扱われるように前処理を行う．

## 3.5 着目条件の指定方法

### 3.5.1 数値データ分析ツールの着目条件

TETDM 上で数値データの明示と絞り込みを行うために用意した処理ツール「データマイニング」を図 4 に示す．

数値データ分析ツール「データマイニング」による着目条件の設定方法には次の 2 種類がある．

1 分析者による着目条件の設定

2 おすすめによる着目条件

1 の分析者による着目条件の設定は，以下の手順で行う．

1) 着目する属性を選択する．

2) スライダーで属性値の閾値を決定する．

3) 閾値以上か閾値以下を選択する．

すなわち，存在する各数値データの中から，1 つの属性を選択して，その属性値の範囲を選択して指定する．明確な条件設定が分析者の頭の中にある場合は，これを用いることで柔軟な条件設定が可能となる．

しかし，どのような条件を入力すべきかわからない分析の初心者や，一通り思いつく条件を設定した後に条件を模索したい場合などに，簡易に条件を入力できるおすすめ着目条件を，表 2 のように用意する．

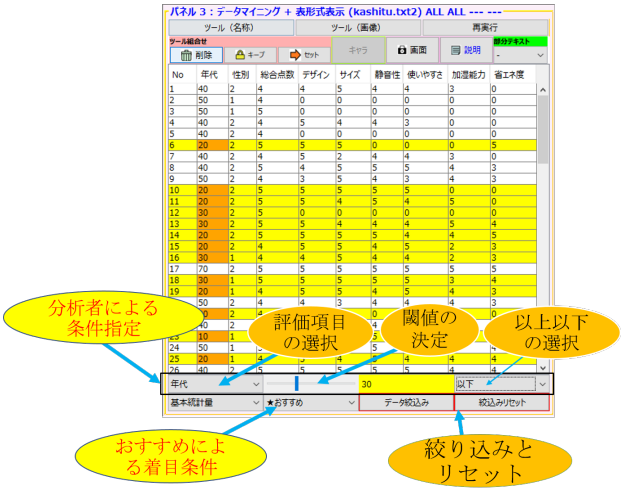


図 4: ツール「データマイニング」の操作画面

表 2: 数値データ分析ツールのおすすめ着目条件

| 条件名 | 条件内容 |
|---|---|
| 最高値 | 各属性で最も高い値を持つもの |
| 最低値 | 各属性で最も低い値を持つもの |
| 最低最高以外 | 各属性で最高値，最低値以外の値を持つもの |
| 最高頻度 | 各属性で頻度が最高の値 |
| 最低頻度 | 各属性で頻度が最低の値 |
| 指定属性値 | 属性と属性値を指定したもの |

すなわち，分析の際に有効となる箇所として，平均値などの基準となる値からのズレを重視して,「最高値」「最低値」「最高頻度」「最低頻度」を用意する．また，データ全体の傾向を探るために,「最高最低以外」「指定属性値」を用意する．なお,「指定属性値」は，与えられるデータに特有の属性と属性値のペアを分析者が指定することを想定している．たとえば，総合評価値が 1 から 5 で表されたデータが入力された場合,「総合評価 1」「総合評価 2」などの条件を設定することで，データの傾向を捉えやすくなると考えられる．

### 3.5.2 テキストデータ分析ツールの着目条件

TETDM 上でテキストデータの明示と絞り込みを行うために用意した処理ツール「単語抽出」を図 5 に示す．

テキストデータ分析ツール「単語抽出」による着目条件の設定方法には次の 3 種類がある．

1 分析者による着目条件の設定

2 おすすめによる着目条件

3 単語指定による着目条件

1 の分析者による着目条件の設定方法は，数値データの場合と同様になるが，設定できる項目は,「単語頻


図 5: 単語抽出（マイニング）の操作画面

表 3: テキストデータ分析ツールのおすすめ着目条件

| | |
|---|---|
| 単語頻度 1 | 文章中で頻度が 1 の単語 |
| 単語頻度最大 | 文章中で頻度が最大の単語 |
| 主語頻度 1 | 文章中で主語としての頻度が 1 の単語 |
| 主語頻度最大 | 文章中で主語として頻度が最大の単語 |
| セグメント頻度 1 | セグメント中で頻度が 1 の単語 |
| セグメント共通語 | セグメント頻度が 2 以上の単語 |
| セグメント頻度最大 | セグメント中で頻度が最大の単語 |
| 100 文字以上 | セグメントの文字数が 100 文字以上のデータ |
| 100 文字以下 | セグメントの文字数が 100 文字以下のデータ |

度」「主語頻度」「セグメント頻度」「文字数」となっている．

また，3 の単語指定による着目条件では，入力された単語とその単語を含むセグメントが明示される．

2 のおすすめの着目条件では，数値データの時と同様に，条件設定に悩んだ時に用いられる条件を，表 3 のように用意する．すなわち，平均値などの基準となる値からのズレを重視した条件として，単語を明示する条件として，各頻度が 1 または最大の単語を条件として用意する．また，データの傾向を捉えやすくするための条件として，セグメント頻度が 2 以上の単語を「セグメント共通語」として，セグメントの長さに着目して，文字数が比較的少ないものとして「100 文字以下」，文字数が比較的多いものとして「100 文字以上」の条件を用意する．

## 3.6 着目点の明示機能

3.5 節にて指定された着目条件に合致するテキスト内の箇所をハイライトにより明示する．

すなわち，条件に合致する数値や単語の背景色をオレンジ色で表示し，それらの数値や単語を含むセグメントの背景色を黄色で表示する．

これによって，条件に合致する部分のみに着目することができ，条件に合致するデータについてのみ，より深い分析を行いたい場合には，次節で述べるデータの絞り込みを行う．

表 4: TETDM 上に用意するツールセットの一覧

| ツールセット名 |
|---|
| 1. データの確認と絞り込み |
| 2. テキスト要約 |
| 3. 失礼単語確認 |
| 4. テキスト分類 |
| 5. 単語頻度確認 |

## 3.7 明示データの絞り込み機能

図 4 や図 5 の「データ絞込み」ボタンを押すことで，前節でハイライトされているデータのみにデータを絞り込むことができる．データの絞り込みを行った後は，データ分析の各ツールは，絞り込まれたデータのみの結果を表示する．

そのため，分析者が与えた条件に合致するデータの特徴を見出しやすくなると考えられる．また，この条件の設定と絞り込みを繰り返し行うことで，より詳細な条件に合致するデータについての分析を行うことが可能となる．

## 3.8 着目点の明示と絞り込み機能を利用したツールセット

本節では，3.5 節で述べたテキスト分析ツール「単語抽出（マイニング）」とデータ分析ツール「データマイニング」を，TETDM の既存のツールと組み合わせた図 4 に示す 5 つのツールセット（TETDM 上のツールの組み合わせとなるパネル構成）について述べる．

### 3.8.1 データ確認と絞り込み

ツール「単語抽出」と「データマイニング」のみを利用した最もシンプルなツールセットとして用意する．

### 3.8.2 テキスト要約

ツール「単語抽出」と「データマイニング」に加え，文章要約のツールを利用できるツールセットを用意す

る．キーワードや重要文を確認することで，重要なことが書かれているデータを探す事に役立てられる．

#### 3.8.3 失礼単語確認

ツール「単語抽出」と「データマイニング」に加え，失礼単語を確認するツールを利用できるツールセットを用意する．失礼な表現や否定的な表現を参照することで，改善点の検討が可能になると考えられる．

#### 3.8.4 テキスト分類

ツール「単語抽出」と「データマイニング」に加え，テキストを分類するツールを利用できるツールセットを用意する．データ全体の傾向を眺めることで，どのようなデータが多いかを確認できると考えられる．

#### 3.8.5 単語頻度確認

ツール「単語抽出」と「データマイニング」に加え，単語の情報を確認するツールを利用できるツールセットを用意する．レビューの中でよく使われている単語を確認し，単語の頻度から着目する単語を発見できると考えられる．

## 4 着目点の明示機能の評価実験

本研究で提案する着目点の明示機能が，データ分析の支援に有効かを検証した実験について述べる．

### 4.1 実験内容

楽天市場みんなのレビュー [7] から「ヒーター」のレビューデータ 150 件と,「加湿器」のレビューデータ 100 件を分析してもらい，レビューの総合点数が高くなると予想される新製品の提案をする目的で，商品のレビューデータ[2] を分析してもらった．

実験は，3.5 節で述べた，おすすめの着目条件を利用可能な提案グループと，利用できない比較グループとに分けて行った．被験者は，理系の大学生，大学院生の 10 名で，各グループ 5 名ずつで実験を行った．

[2]本レビューデータは，レビュアの年齢や性別および，商品の評価が 5 段階評価されている数値データとレビューのテキストデータがセットになっている．

表 5: 集められた解釈の数（被験者平均）

| | ヒーター | 加湿器 |
|---|---|---|
| 提案グループ | 10.3 | 10.5 |
| 比較グループ | 7.5 | 8.6 |
| 差 | 2.8 | 1.9 |

#### 4.1.1 実験手順

以下に被験者に提示した，評価実験の手順を示す．

1. 分析に用いるツールセットを選択する．
2. 着目点の明示機能によりハイライトされた数値や単語からデータを絞り込む．
3. 絞り込んだデータの特徴を解釈する．もしくは，さらにデータを絞り込む．
4. 手順 1 から手順 3 を繰り返して解釈をできるだけ多く登録する[3]．
5. TETDM の知識創発機能[4]を用いて，集めた解釈の共通点を見出すことで解釈をまとめる．
6. 共通点が見つからない段階まで解釈をまとめてもらい, それを最終提案とする．
7. 最終的に共通点を絞り込みきれなかった場合は，まとめた解釈を並べて複数文を最終提案とする．

### 4.2 実験結果と考察

#### 4.2.1 「結果と解釈」の登録数

表 5 に，おすすめの着目条件を利用できる提案グループと，利用できない比較グループの被験者が集めた解釈の数を示す．ただし，同一内容の解釈は 1 つとしてカウントしている．

「ヒーター」「加湿器」のいずれのレビューについても，提案グループの方がより多くの解釈を集めることができていた．これは，分析者が自分で思い描く条件だけで分析するよりも，分析者の頭の中にはない条件を提示することで，より多くの解釈を導くことができたためと考えられる．また，分析者が自分で条件を設定するためには，項目の選択，閾値の選択，上限か下限の設定，の 3 つのステップを経る必要があるのに対して，おすすめの条件設定においては，プルダウンメニューの中から 1 つの条件を選択すればよかったため，

[3]TETDM にある「結果と解釈」の登録機能を用いて，データからわかったことを記録してもらう．

[4]登録された結果と解釈を統合して 1 つにまとめるための支援インタフェース

表 6: 最終提案の着目点の数と具体性（被験者平均）

| | 着目点 | 具体性 |
|---|---|---|
| 提案グループ | 3.0 | 2.9 |
| 比較グループ | 2.0 | 1.6 |
| 差 | 1.0 | 1.5 |

表 7: ヒーターの最終提案（提案グループ）

| 回答者 | 最終提案 | 着目点 | 具体性 |
|---|---|---|---|
| 提案 A | 年齢，性別を通して小さく，パワフルであることが利点であり，発送トラブルやデザインの誤解が評価を下げてしまう | 5 | 2 |
| 提案 B | 年齢に関しては，お年寄りは利便性のみに関心が強く，若者は利便性以外にも発送時間に不満がある人が多かった．また，性別のに関しては，女性にとって，デザインとサイズが高評価であった． | 8 | 2 |

表 8: ヒーターの最終提案と評価（比較グループ）

| 回答者 | 最終提案 | 着目点 | 具体性 |
|---|---|---|---|
| 比較 A | この商品はユーザが求める「コンパクトさ」を満たしていると考えられるが，さらにコンパクトさを追求する必要がある | 0 | 1 |
| 比較 B | サイズがコンパクトであり，すぐに部屋が暖かくなること | 1 | 0 |

そのような条件設定の簡易さも，この差につながった可能性があると考えられる．

#### 4.2.2 最終提案の評価の比較

集めた解釈をもとに知識創発を行い，まとめてもらった最終提案についての評価を行った．評価は，データの絞り込みに用いられる属性について，提案の中で言及している属性の数[5] と，具体性として，商品の長所や改善点が具体的に書かれていると判断できる提案[6]を数えた．

結果を表 6 に示す．着目点と具体性のいずれも提案グループの値の方が大きい結果となった．このことから，おすすめの着目条件を用いた被験者の方が，より幅広く具体的な提案につなげることができたことがわかる．

実際の提案の例を表 7 と表 8 に示す．提案グループの被験者の方が，集められている解釈の数が多かったため，それらを用いた幅広い視点からの提案につなげることができたと考えられる．

## 5 結論

本研究では，データにおける平均などの基準値からのズレが大きいデータを，おすすめの着目条件として選択，明示できるようにするとともに，与えた着目点を手がかりとして，データへの着目と絞り込みを繰り返し行える機能を実装した．

評価実験により，おすすめの着目条件が幅広く具体的な考察を行うために有効なことを確認した．

今後は，より効果的なおすすめ条件設定として，複数の属性に関わるおすすめ条件や，データの分布に依存したおすすめ条件の設定を検討していきたいと考えている．

[5] 年齢や性別，評価点数やデザインサイズといった項目に加え，年齢層を示す「若者」や加湿（加熱）能力を示す「パワフル」といった単語も着目点としてカウントしている．

[6] 「インテリアに使える」や「センサーの感度向上」のように，長所や改善点の，方向性や程度がわかる場合にカウントする．

## 参考文献

[1] 赤峯享：ビッグデータ分析でのテキスト情報の活用, 自然言語処理, Vol.20, No.5, p.627, (2013).

[2] 砂山渡, 高間康史, 徳永秀和, 串間宗夫, 西村和則, 松下光範, 北村侑也:統合環境 TETDM を用いた社会実践, 人工知能学会論文誌, Vol.32, No.1, NFC-A, pp.1–12, (2017).

[3] 松尾哉太, 新妻弘崇, 太田学：レビュー解析に基づくユーザ評価の根拠提示の一手法, 情報処理学会研究報告, Vol.35, No.14, pp.1–6, (2014).

[4] 稲垣和人, 吉川大弘, 古橋武：アンケートデータ解析におけるマイノリティの抽出手法に関する検討, 日本知能情報ファジィ学会第 76 回全国大会講演論文集, No.1, pp.383–384, (2014).

[5] 越康治, 高田淑子, 木下英俊, 安藤明伸, 高橋潔, 田幡憲一, 岡正明, 石澤公明:テキストマイニングによる授業評価アンケートの分析:共起ネットワークによる自由記述の可視化の試み, 宮城教育大学情報処理センター研究紀要:COMMUE, No.22, pp.67–74, (2015).

[6] 松本友哉, 砂山渡, 畑中裕司, 小郷原一智：データマイニングとテキストマイニングの連携によるデータ分析支援, 第 15 回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会資料, pp.14–19, (2017).

[7] 楽天みんなのレビュー：
(URL) https://review.rakuten.co.jp/