既存データセットの活用事例に基づく新規データセットの活用方法推薦に関する検討

林　沙也加†　　畑　　玲音†　　松下　光範††

† 関西大学大学院総合情報学研究科　〒 569–1095 大阪府高槻市霊仙寺町 2–1–1
†† 関西大学総合情報学部　〒 569–1095 大阪府高槻市霊仙寺町 2–1–1
E-mail: †{k918726,k223167,t080164}@kansai-u.ac.jp

あらまし　本研究の目的はユーザのデータ利活用を支援することである．現在提供しているサービスの有用性検証や新規サービスの考案など，様々な利用目的の下でユーザの購買データやレビューデータなど様々な大規模データセットが日々構築されている．こうしたデータセットを分析して有益な情報を抽出するために様々な技術が開発されているが，ユーザが目的に沿った情報を取得するには，適切な技術の選択が不可欠である．しかし，技術に不慣れなユーザが未知のデータに対して適用すべき技術を適切に選択することは容易ではない．そこで本稿では，データの特性，適用する分析技術，得られる結果の三者の関係に注目し，既存の活用事例を分類整理することで，ユーザが未知のデータを扱う際の支援を試みる．提案手法では，類似した特性を持つデータには同じ分析技術が適用可能であるという仮説のもと，データセット内のカラムの形式と性質の類似性から適用可能な技術を推薦する．

キーワード　データセット活用支援、データ特性、データの類似性

1　はじめに

商品目録データや販売データなど，企業活動の際に蓄積される大量のデータを整理し，大規模データセットとして活用したり共有したりする事例が散見されるようになってきている．こうした大規模データセットには，ユーザの行動やニーズを推測する上で手がかりとなる情報が記録されており，将来のビジネス活動の貴重な情報源となっている．例えば，EC サイトを運営する企業では，サービスの運用過程でユーザの購買データやレビューといったデータが日々収集・蓄積されており，これらのデータを大規模データセットとして整備し分析することで，既存サービスの有用性を検証したり，新規サービスを考案する際の手がかりにしたりすることができるようになる．

こうしたデータセットを分析し，有益な情報を抽出するために，多様な技術が開発されている．例えば，不動産情報サイトに掲載された物件データをもとに場所や間取りを考慮した物件の価格相場を算出するアルゴリズムの開発 [11] や携帯端末の位置情報を利用し，旅行者が興味を持つ対象を推定する技術 [1] などが挙げられる．

ユーザの保有するデータセットを利用目的に沿って活用するためには，蓄積されたデータ（e.g., 識別 ID や評価値などの数値データ，説明文や評価文などのテキストデータ）の特性やデータ構造を考慮し，適切な技術を選択することが不可欠である．しかし，データの活用方法に対する知識が乏しかったりデータ処理技術に不案内だったりするユーザの場合，大量のデータの内容やそのデータ特性を十分に把握することができず，目的に合った分析手法や技術を適切に選定できないという問題がある．また，既存のデータ活用支援システムは，限られた条件に基づく過去の事例に依存していることが多く，新規のデータセットや多様な分析目的に対応することが難しいという課題が存在する．そのため，データ特性を考慮し，新規のデータセットにも柔軟に対応可能なデータ活用技術の推薦手法が求められている．

本研究では，データ活用事例において，データの特性（e.g., レビューデータ，評価データ），適用する分析技術（e.g., BERT, SVM，Topic Modeling），得られる結果（e.g., カテゴリー分類，評価値推定）の三者の関係性に着目し，既存の事例を整理・分類することにより，新規データセットに適用可能な技術を推薦する手法を提案する．提案手法では，「類似する特性を持つデータには同じ技術が適用可能である」という仮説の下，データセット内のカラムの形式と性質の類似性から適用可能な技術の推薦を行う．その端緒として，本稿ではデータセット内のカラムの形式と性質に基づくデータセット間の類似性の評価を試みる．これにより，ユーザが保有するデータセットの特性と利用目的に基づいて適切な技術を選定するための支援を行い，データ利活用の意思決定を迅速かつ効果的に支援することを目指す．

2　関連研究

データセットの利活用を支援するために，データセットの研究利用を対象とした研究がいくつか報告されている．本章では，データセットの情報表現について概観し，本研究の位置づけを明らかにする．

早矢仕らは，データ自体を共有することなく，その概要情報を基にデータの価値を評価できる「データジャケット」と過去のデータ利用案（要求・ソリューション）を結びつけることで，ユーザが目的に合ったデータを検索できるシステム「Data Jacket Store (DJ Store)」を構築した [8]．このシステム

では RDF (Resource Description Framework) を用い，要求・ソリューション・データジャケット間の関係をモデル化し，構造的に表現する手法を採用している．これにより，単純なキーワード一致だけでなく，過去のソリューションや要求を介して関連するデータジャケットを発見することが可能となった．しかし，このシステムの実験結果は，限られたデータセットと過去の要求・ソリューションに依存しており，新たな領域や未知のデータ利用には必ずしも対応しきれないという課題があった．また，データジャケットはデータ全体の概要を把握するためのものであり，個々のデータカラムの詳細情報を扱うものではないため，ユーザの目的に沿ったデータカラムを特定することが容易ではないという問題が指摘されている．

玄道らは，コンテンツとそのコンテンツに関わる人（クリエイター・ユーザ）の関係に基づき，複数のデータセットを抽象化し整理するフレームワークを提案した [3]．このフレームワークでは，データ項目を「コンテンツ」,「クリエイター」,「ユーザ」の三者の観点から表現し，それぞれのデータ項目間の関係性を RDF 形式で抽象化している．RDF 形式における述語として，目的を説明する「describe」，目的を評価する「evaluate」，目的を創造する「create」，目的を使用する「use」，目的に返答する「reply」の 5 種類を選定している．また，異なるデータセット間で同じ意味を持つデータ項目（e.g., 楽天市場データセットの「レビュー内容」と楽天トラベルデータセットの「ユーザ投稿本文」）を共通項目として整理し，そのデータセットを利用した研究に関する論文同士の類似性を検証している．フレームワークを用いることで，異なるデータセット間の共通項目をもとにした類似性を見出すことができたことから，提案フレームワークの有効性が示されたとしている．しかし，選好（評価値）を離散値で表現する場合（e.g., [1,2,3]）とテキスト形式で表現する場合（e.g., [悪い，普通，良い]）があるように，一部のデータ項目では同じ役割で用いられているにもかかわらず表現形式が異なる場合があり，完全な共通化が困難であることや，共通項目に該当しないデータ項目の関係性を理解するには限界があり，全てのデータ項目に対応することはできていない．また，このフレームワークはユーザが実際にデータ項目を使用する場面（e.g.,「レビュー内容」に TF-IDF を適用して重要単語を算出する）を十分に考慮していない点も指摘されている．

そこで本研究では，これらの課題を解決するために「データ特性」「適用する分析技術」「得られる結果」の三者の関係に着目し，既存の活用事例を整理・分類することで，未知のデータセットに対して適切な分析技術を提案するための枠組みを構築する．これにより，データセットの特性とそれに適用する技術の組み合わせを自動で抽出可能にすることで，未知のデータセットに対して適用可能な技術を判定できる．また，既存の活用事例はデータセット活用の成功事例となるため，ユーザがデータセットの適切な分析方法を把握可能になる．

本稿では，データセットの類似度を推定する手法を提案する．活用事例で活用されたデータセットと未知のデータセットの性質が類似していると活用事例で適用された技術を未知のデータセットに適用できる可能性がある．この手法により，既存事例のデータセットと類似度が高いと推定されたデータセットに対して，比較元のデータセットで適用された技術を適用することで，比較元の結果と類似した結果が得られることが期待される．この検証を通じて，提案する枠組みの妥当性を評価する．

3 提 案 手 法

本稿では，既存のデータセット活用事例をもとに，データ特性，適用する分析技術，得られる結果の三者に着目し，既存の活用事例を「データ × 技術 × 結果」の形式で整理する手法を提案する．既存の活用事例として研究論文を採用し，論文に対する考察から，論文内では使用するデータセット，それに適用した分析技術，得られた結果の三者が記述されている事例が多いことがわかった．そのため，「データ × 技術 × 結果」の形式を採用することで，データ，技術，結果のいずれか二つの情報が揃っている場合に，欠けているもう一つの情報を推定し，ユーザが未知のデータの利活用にも対応できる仕組みの構築を試みる．

1 章で述べたように，企業や研究者は日々新たなデータを収集しており，既存の活用事例で使用されたデータセットとは異なる特性を持つことが多い．また，新しいビジネス要件や研究目的を達成するために，ユーザの目的に沿った分析技術が必要な場合がある．未知のデータやユーザの目的に対して適切な分析技術を選定することはデータの特性を十分に把握していないユーザにとって困難であるため，これらに対応可能な分析手法を推薦する仕組みが求められている．こうした背景を踏まえ，本稿では「データ × 技術 × 結果」の形式を採用することで，未知のデータセットや新しい分析目的への対応を支援することを目指す．例えば，以下のようなケースに対応できることが期待される．

未知のデータセットの場合

ユーザが持つ未知のデータセットが既存の活用事例で使用されたデータセットに類似している場合，活用事例の「技術」と「結果」を推薦できる．これにより，ユーザがデータ特性に基づいて適切な技術を選定することが可能となる．
例）ユーザが「皿」のレビューデータ（未知）を保有しており，服のレビューデータ（既存）と特性が類似している場合，服のデータに適用された技術と結果を皿のレビューデータに応用できる．

ユーザの目的が明確な場合

ユーザが目的を明確に持っていても，技術に対する知識や経験が乏しい場合，適切な技術や必要なデータが判断できない場合がある．この際，ユーザの目的と類似している活用事例の「結果」を推薦に使用する．ユーザの目的が明確な場合，ユーザがデータを所持している場合と所持していない場合に細分化される．

データを所持している場合

ユーザの所持しているデータが既存のものか未知のも

表 1: 「データ × 技術 × 結果」の形式でまとめた論文

論文名	データ	技術	結果
擬似コーパスを用いた 飲食店レビューの観点の自動分類 [12]	クックパッドのつくれぽと楽天トラベルのレビュー	fastText	レビュー観点の分類
単語の頻度と意味に基づいたコミックに関する テキスト情報源の特性分析 [10]	コミック作品のキャラクタ説明文と評価サイトのレビュー文	クラスタリング	各特徴の明確化
化粧品の評価項目別スコア生成のための 評価表現辞書の自動構築 [5]	化粧品レビューと評価値	Word2Vec	評価表現辞書の構築
半教師あり NMF を用いた 専門分野と講義の関係推定 [13]	卒業論文概要とシラバス	半教師あり NMF	専門分野と講義の関係性の可視化
コミックの登場人物についての 説明文からの性格タグ推定 [9]	キャラクタ説明文と性格タグ	ニューラルネットワーク	性格タグを推定
EC サイトのレビューテキストからの レーティング予測と購買者評価の分析 [4]	評価値とレビュー	BERT リッジ回帰 RNN	評価値の予測
レビュースコアとレビューの文字数に着目した 商品スコアの再計算 [6]	商品レビュー	文字数のカウント	評価値の予測
過賞賛のレビューの検出と特徴語抽出 [2]	評価値とレビュー	SVM	過賞賛レビューの検出

のに細分化される．

- ユーザが所持しているデータが既存の場合，既存のデータと結果から技術を把握できる．
  例）ユーザが服のレビューデータ（既存）を保有しており，目的が明確であれば，服のデータに適用されている技術を推薦できる．
- ユーザが所持しているデータが未知の場合，未知のデータと類似している既存のデータを推定し，結果と組み合わせることで技術を把握できる．
  例）ユーザが「皿」のレビューデータ（未知）を保有しており，皿の組み合わせを自動で選定するという目的がある場合，皿データと類似しているデータとして服のレビューデータ（既存）が推定される．「服データ」と「服の組み合わせを自動で選定する」という目的から服のデータに適用された技術をユーザに推薦する．

データを所持していない場合

ユーザの目的に沿った結果を達成するために必要なデータの特性や技術をユーザが把握できる．

例）ユーザがデータを保有していない場合，「組み合わせを自動で選定する」という目的と類似した既存の事例から服のレビューデータが目的を達成するのに有効なデータ特性であることがわかる．これによって，ユーザは目的を達成するために必要なデータの特性を把握することができる．

既存の活用事例を「データ × 技術 × 結果」の形式で整理した例として，レビューテキストと評価値から BERT を用いてレビューテキストからユーザがつけた評価値を予測するといった活用事例 [4] では，データとして「評価値とレビュー」，技術として「BERT」，結果として「評価値の予測」が挙げられる．この場合，レビュー文が持つデータ特性（e.g., テキスト形式，評価要素）に基づき，BERT が適用可能であることが示される．同様に，時系列のエネルギー需要データから LSTM を用いてエネルギーの需要を予測するといった活用事例 [7] では，データとして「時系列データ」，技術として「LSTM」，結果として「需要の予測」が整理される．この場合，時系列データの特性（連続的な構造）を考慮し，LSTM による需要変動の予測が適用可能であることが示される．上記の形式で整理した論文を表 1 に示す．

このように，「データ × 技術 × 結果」の形式を用いて活用事例を整理することで，技術に不慣れなユーザでも保有するデータセットの特性と得たい結果に基づいた，適切な技術を選定する際の指針を得ることができる．

4 データセットの類似度

「データ × 技術 × 結果」の関係性に基づいた未知データへの活用手法が成り立つという仮定のもと，データセット間の類似度を測定する手法を提案する．この類似度測定により，既存の活用事例が未知のデータセットに適用可能かを検証し，技術の選定や結果の予測に役立つかについて検証する．その際，データセットや技術，結果の類似度の算出には，データカラムとそれに付与されたメタデータを管理するデータベース（以下，データ DB と記す．），分析技術を管理するデータベース（以下，技術 DB と記す．），および分析結果を管理するデータベース（以下，結果 DB と記す．）を活用する．

4.1 「データ × 技術 × 結果」の形式に整理する手法

「データ × 技術 × 結果」の形式でユーザに技術や結果を推薦する際の流れを図 1 に整理する．まず，ユーザの保有するデータセット，ユーザが使用可能な分析技術，およびユーザが得たい結果を入力とする．この入力からの処理の流れは以下の通りである．

（1） ユーザの保有するデータセットの各データカラムに対して，データ DB 内の類似データカラムを探索し，そのメタデータを取得する．

（2） 取得したメタデータと類似するメタデータを持つデータカラムをデータ DB から検索し，類似カラムとして取得する．

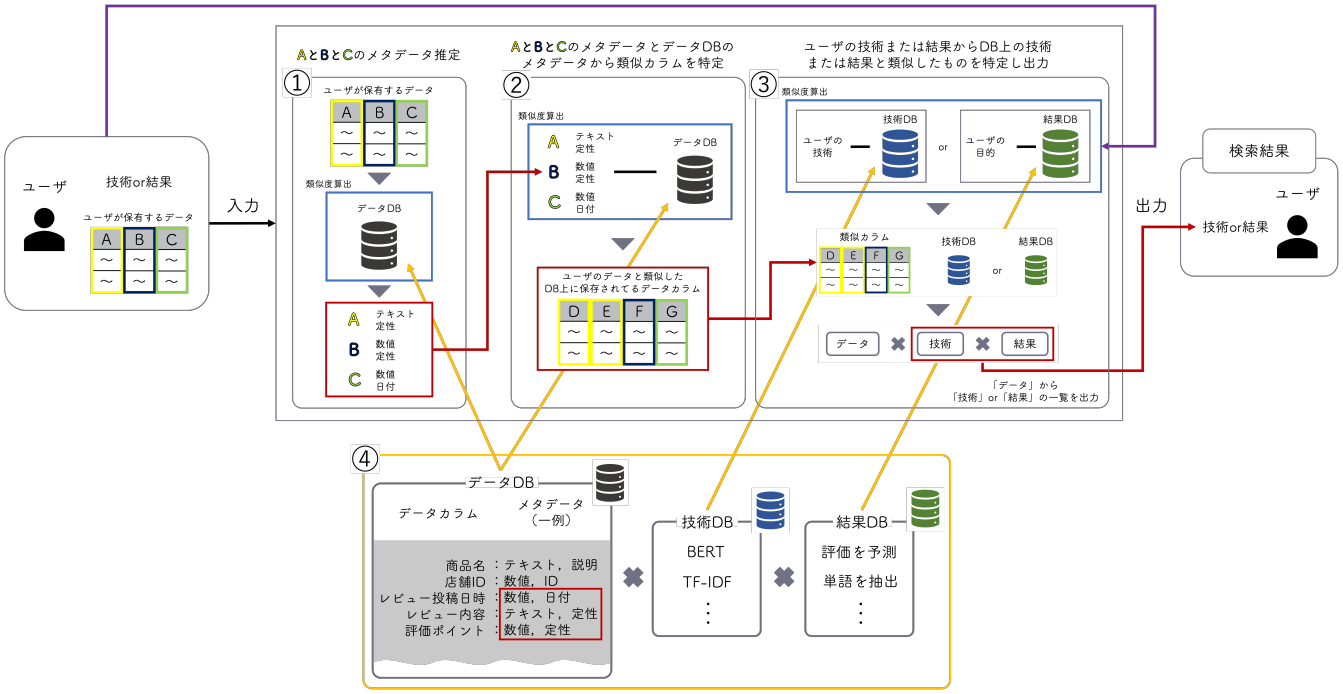

図 1: ユーザの新規データ利活用の図

（3） ユーザが使用可能な技術と類似する技術を技術 DB から検索する（類似技術）．または，ユーザが得たい結果と類似する結果を結果 DB から検索する（類似結果）．

（4） 最後に，取得した類似カラム，類似技術，または類似結果を「データ × 技術 × 結果」として統合し，適切な技術または結果をユーザに提示する．

本手法では，図 1-④のデータ DB，技術 DB，結果 DB から，図 1-①，②，③の類似度推定を行うことで活用事例を整理するため，データ，技術，結果それぞれの類似度を測る必要がある．その一端として，本研究ではデータの類似度を推定する手法を提案する．

4.2 データセットの類似度推定手法

「類似した特性を持つデータには同じ分析技術が適用可能である」という仮説のもと，データセットにメタデータを付与し，そのメタデータを用いて類似度を算出する手法を提案する．メタデータは，データセットの形式や性質を表すラベルとして付与される．山西らは飲食店のレビュー文中にホスピタリティと料理の観点が混在していることに着目し，このレビュー文を機械的に分類するために，ホスピタリティに関する特徴を持つホテルレビューコーパスと，料理に関する特徴を持つレシピレビューコーパスを組み合わせた擬似コーパスを用いて分類器を作成している [12]．この研究は，データの作成された観点（料理について書かれたレビューやホスピタリティについて書かれたレビュー）と同様の観点を持ち合わせたデータセットであれば飲食店レビューの分類器を作成できるといったものであるが，同様の観点があるかどうかは著者が感覚的に行っていると推測される．そこで，類似しているデータセットを特定するために，各データセットに対して形式と性質の観点からメタデータを付与した．

本研究では，どのような特性を持つデータセットが活用事例で使用された分析技術に適用可能であるかを推測し，その特性を持つデータセット同士は類似していると定義する．例えば，「楽天市場データセットのレビュー文と評価値を BERT で分析して，未知のレビューデータに対して評価値を推測する」といった事例では，BERT を実行するためにデータセットに必要な特性を推測する．ユーザが「評価する」といった性質を持つテキストデータと，レビューデータから評価値を予測するための正解データとして実際にユーザが付与したカテゴリである評価値のデータが必要と考えられる．そのため，BERT を用いて未知のレビューテキストから評価値を推測するには，「文章である，かつ，ユーザが評価している」という性質を持つテキストデータと，「順序尺度に当てはまる，かつ，ユーザが評価している」という性質を持つ数値データを使用することが好ましい．類似したデータセットが複数存在した場合，付与したメタデータが一致している数が多いものほど類似度が高く，少ないものほど類似度が低いとした．これは，メタデータはデータセットの特性を表すものであり，データセット同士でメタデータの一致する数が多いものはデータの特性が一致していることになると考えたからである．この手法により，類似したデータを特定し，既存の活用事例をもとに未知のデータに適用可能な技術を推定する．

4.2.1 メタデータの付与

本研究では，データセットの特徴をメタデータを用いて表す際，データの表層的なものとしてデータの形式的特徴（e.g., テキスト，数値），深層的なものとしてデータの性質的特徴（e.g., 名義尺度，順序尺度）を考慮する．例えば，評価レビュー（テキスト）と評価値（数値）を使用する場合，どちらも商品など

の評価をしている点では共通しているが，評価レビューはテキスト形式，評価値は数値形式とデータの形式が異なる．そのため，評価レビューは評価したユーザの使用語彙の傾向を知るといった利用ができるが，評価値はできない．また，評価値（数値）と商品価格のデータ（数値）を使用する場合，どちらも数値形式であるという点では共通しているが，日付のデータは順序尺度，商品価格のデータは比例尺度とデータの性質が異なる．そのため，商品価格のデータは「1000 円から 2 割引で 800 円になる」というような比率計算ができるが，評価値は数字間の差が不均等なためそのような計算ができない．これらの理由から，各データセットに形式的特徴と性質的特徴の二種類のメタデータを付与する．例えば，数値データは「名義尺度」や「順序尺度」に，テキストデータも「評価文」や「説明文」に分類することができる．これにより，データセット同士の類似度を形式的および性質的観点から評価することが可能になることから，それぞれの類似度を測定することが可能である．さらに，形式や性質が異なる場合でも，類似性を見出せるケースが存在する．例えば，テキストデータの「多い・普通・少ない」といった表現は，テキストであるが意味的には「順序」の性質をもっていると考えられるため，数値データの「順序尺度」と同等に扱うことができる．このような形式的および性質的なデータ特性をもとにデータセット間の関係性を評価することが，提案手法の重要な基盤となる．

また，データセットには複数のカラムが含まれており，ユーザがデータセットを利用する際には，カラム単位でデータを扱うことが一般的である．そのため，提案手法では，各カラムに対して形式や性質を表すメタデータを付与することで，データセット全体の構造をより明確化した．付与されたメタデータは，データカラム同士の類似性評価を可能にし，適切な分析技術の推薦に活用することが期待される．

付与したメタデータは階層構造を持ち，データの形式的特徴と性質的特徴を包括的に表現する．テキストや数値の特徴は細分化することができる．例えば，テキストであれば単語やテキストに，数値であれば順序尺度や間隔尺度に細分化して，階層構造を持つメタデータを付与した．メタデータの階層構造を図 2 に示す．

上記の議論を踏まえ，本研究では，各データカラムに設定するメタデータを (1) 形式に関する観点，(2) 内容に関する観点，の 2 つの観点で整理することとした．

形式に関するメタデータは以下の通りである．

- **テキスト**：数値以外のデータ
  - **文章**：テキストが文章として成り立つデータ（e.g., レビュー内容）
  - **単語**：テキストが文章として成り立たないデータ（e.g., 単一のキーワード，タグ）
- **数値**：数値のみのデータ
  - **名義尺度**：グループ分類を示すデータ（e.g., カテゴリラベル）数字の順序や大きさに意味はない
  - **順序尺度**：順序を表すデータ（e.g., ランク）順序間の差は一定ではない
  - **間隔尺度**：差が一定のデータ（e.g., 日付，温度）比率は計算できない
  - **比例尺度**：差が一定で比率計算が可能なデータ（e.g., 重量，金額）

内容に関するメタデータは以下の通りである．

- **記述データ**：データそのものを説明する記述についての情報
  - **名前**：名前やタイトル（e.g., 店舗名，素材名）
  - **目的**：用途や対象（e.g., 商品の使い道，誰のために購入したか）
  - **説明**：対象の説明（e.g., 商品・施設の詳細説明）
  - **カテゴリ**：分類や種類（e.g., 商品のカテゴリ）
  - **手続き**：手順の順序を表すデータ（e.g., レシピ手順，作業手順）
- **指標データ**：他のデータとの比較など，評価（e.g., 優劣，差異）の指標として用いることが可能なデータ
  - **定量**：数値として扱うことができるデータ
    - **回数**：利用回数や訪問回数の回数を表すデータ
    - **日付**：日時を表すデータ（e.g.,「春分の日」などのテキストで表された日付）
    - **値段**：金額を表すデータ（e.g., 商品の価格，施設料金）
    - **量**：数量や容量（e.g., 調味料の量，商品数）
    - **順序**：順序を表すデータ（e.g.,「良い」「普通」「悪い」や「とても満足」「どちらでもない」「とても不満」）
  - **定性**：評価や感想（e.g., 評価点数，レビューテキスト）
- **その他**：上記のカテゴリに該当しない特殊なデータ
  - **ID**：データを一意に識別するための情報（e.g., ユーザ ID，商品 ID）
  - **URL**：データリソースの所在を表すデータ（e.g., 商品の HP の url）
  - **パス**：データリソースのパスを表す情報（e.g., 商品カテゴリの根ノードまでのパス）

4.3 実装手順

本稿では，4.2 節で定義したメタデータを楽天データセット [14] の中から楽天市場データセット，楽天トラベルデータセット，楽天レシピデータセット，および楽天 GORA データセットに付与した．これらのデータセットを使用した理由として，データ量が比較的多く，メタデータ付与や解析を行う際に十分な規模を有していること，研究においてさまざまな用途での利用がされており，提案手法を評価するために適切と判断したためである．使用したデータカラムの種類を表 2 に示す．

データカラムごとにメタデータを付与する手順は以下の通りである．

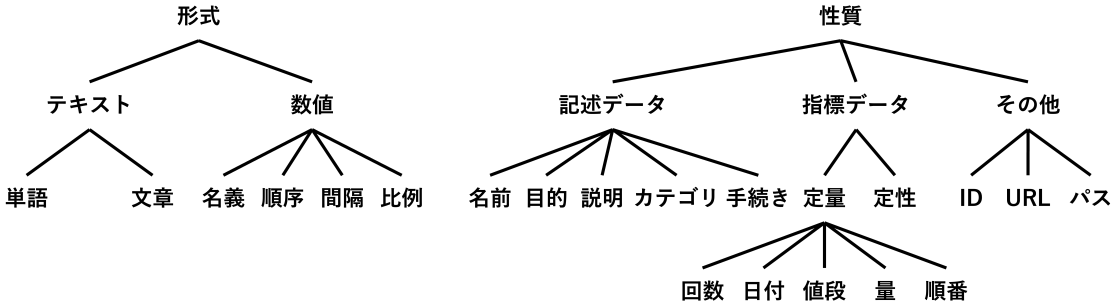

図 2: メタデータの構造

表 2: メタデータを付与したデータセット名とデータカラム名の一例

商品データ	商品名，店舗コード，商品コード， 商品ページ URL，商品価格，商品ジャンル ID， 商品画像 URL，販売方法別説明文，商品説明文
ジャンルマスタ	ジャンル ID，親ジャンル ID，ジャンル名
みんなのレビュー・口コミ情報	投稿者 ID，店舗名，店舗 ID，商品名，商品 ID， 商品ページ URL，商品ジャンル ID， 商品ジャンル ID パス，使い道，目的，頻度， 評価ポイント，レビュータイトル，レビュー内容， 参考になった数，レビュー登録日時
市場店舗レビューデータ	投稿者 ID，店舗名，店舗 ID， 評価ポイント，レビュー本文， 参考になった数，レビュー投稿日時

（1） 各データカラムの形式（e.g., テキスト，数値）を確認する．

（2） テキストの場合，単語か文章かを区別する．数値の場合，名義尺度，順序尺度，間隔尺度，比例尺度のいずれに該当するかを判断する．

（3） カラムが「コンテンツ」（e.g., 名前，説明，目的）や「比較可能性」（e.g., 評価，回数，日時）に該当するかを検討する，

（4） 上記の判断に基づき形式・性質をメタデータとして記録する．各項目に「1」（該当する）または「0」（該当しない）を付与する形で数値化する．

例えば，「234515」（店舗 ID）というデータと「もう何度目のリピでしょうか。 ちゃんと辛いし、エビの旨味もいいし、常備してます。」（レビュー内容）というデータに対してメタデータを付与する．（1）の手順で「234515」に対して「数値」のメタデータを付与し，「もう何度目のリピでしょうか。 ちゃんと辛いし、エビの旨味もいいし、常備してます。」に対して「テキスト」のメタデータを付与する．（2）の手順で「234515」に対して「名義」のメタデータを付与し，「もう何度目のリピでしょうか。 ちゃんと辛いし、エビの旨味もいいし、常備してます。」に対して「文章」のメタデータを付与する．（3）の手順で「234515」に対して「その他」と「ID」のメタデータを付与し，「もう何度目のリピでしょうか。 ちゃんと辛いし、エビの旨味もいいし、常備してます。」に対して「比較可能」と「評価」のメタデータを付与する．このような手順でメタデータを付与した結果として，各データカラムの特性が数値化され，表 3 に示す通り，データセットカラムごとに特徴が整理された．

4.4 類似度の算出方法

楽天データセットを使用している論文として小林らが行った研究から BERT を用いて状況を想定する [4]．楽天市場データセット中の「みんなのレビュー・口コミ情報」から「評価ポイント」と「レビュー内容」のデータカラムを使用している．4.2 節での定義をもとに論文で使用しているデータと類似しているデータカラムを抽出した．表 3 に示したようにメタデータを付与したデータカラムから，「テキスト（文章）」と「指標データ（定性）」の性質（表 3 中の 4，18 番目のカラムに「1」のデータが入っているデータカラム）を持つデータと「数値（順序）」と「指標データ（定性）」の性質（表 3 中の 6，18 番目のカラムに「1」のデータが入っているデータカラム）を持つデータが BERT を実行する上で使用可能なデータであると言える．「評価ポイント」と「レビュー内容」のデータカラムと類似していると判断したデータカラムを以下に記す．

- 楽天市場データセットの中の「みんなのレビュー・口コミ情報」から「評価ポイント」と「レビュータイトル」のデータカラム
- 楽天トラベルデータセットの中の「ユーザ評価」から「ユーザ評価の評価 7（総合）」と「ユーザ投稿本文」のデータカラム
- 楽天 GORA データセットの中の「クチコミ情報」から「総合評価」と「コメント」のデータカラム

5 評価実験

提案手法により算出されたデータカラムの類似度の正当性を評価する．「データ × 技術 × 結果」の関係が正しいという仮定のもと，提案手法により類似していると判断されたデータと類似してないと判断されたデータを使って，既存の活用事例の「データセット」の部分に際のモデルの精度を評価する．

使用する分析技術について，既存の研究成果で使用されている「データ × 技術 × 結果」の組み合わせには正当性があるとして分析技術を採用した．また，本研究では各分析技術の妥当

表 3: 「みんなのレビュー・口コミ情報」に付与したメタデータ

カラム名	テキスト	数値	単語	文章	名義	順序	間隔	比例	記述	指標	その他	名前	目的	説明	カテゴリ	手続き	定量	定性	ID	URL	パス	回数	日付	値段	量	順番
投稿者 ID	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
店舗名	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
店舗 ID	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
商品名	1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
商品 ID	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
商品ページ URL	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
商品ジャンル ID	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
商品ジャンル ID パス	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
使い道	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
目的	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
頻度	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
評価ポイント	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
レビュータイトル	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
レビュー内容	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
参考になった数	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
レビュー登録日時	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0

性を検証することを目的とはしていない．既存の技術を前提として，提案手法によるデータセットの類似性評価が分析技術の適用可能性をどの程度正確に示すことができるかを検討した．

5.1 実験手順

小林らは，EC サイトにおいて購入者のレビューがユーザの商品の評価や購買行動に大きな影響を与えることに着目して，レビューテキストからレーティング（星の数）を予測するモデルを構築し，レビューテキストとレーティングの関係を分析した [4]．データセットは楽天市場の 60,000 件のレビューを使用しており，各レビューには，テキストと 5 段階の評価が含まれている．使用した技術は，ベースラインモデルとして線形回帰，RNN（Recurrent Neural Network），および LSTM（Long Short-Term Memory）を使用し，より高度な手法として Transformer モデルおよび BERT（Bidirectional Encoder Representations from Transformers）を採用し，それぞれの予測精度を比較した．実験の結果，Transformer モデルが最も高精度な予測を実現することが確認された．

小林らが行った研究の構造に基づいた再現によって，実験を行った．小林らは「評価ポイント」と「レビュー内容」のデータカラムの商品カテゴリを「白米」に限定しているが，本研究では楽天市場と同様のデータのみを扱うわけではないため，「白米」と限定すると他のデータセットでの分析技術の再現ができなくなる．そのため，カテゴリは指定せず全てのカテゴリを使用して再現した．BERT の事前学習済モデルには東北大学乾研究室が公開している，「BERT の日本語事前学習モデル」[1]を用いている．実測値（楽天データの評価ポイント）と BERT での予測値との誤差を評価する指標として，二乗平均平方根誤差（RMSE）と平均絶対値誤差（MAE）を用いているため，本研究でも同様の評価指標を用いる．これらの評価指標を用いた理由としては，小林らの研究において，BERT 以外にも RNN やリッジ回帰が使用されており，これらの評価指標として RMSE や MAE が適していたと推測されるためである．本稿では，分析技術として BERT のみを採用しているため，RMSE と MAE に加えて，正解率，適合率，再現率，F 値の値も算出した．

小林らの研究を再現するために，4.4 で推定したデータカラムを BERT を用いて分析した．データカラムをそれぞれ 10,000 件を用意し，学習データとして 8,000 件，テストデータとして 2,000 件に分割した．BERT の事前学習済モデルには東北大学乾研究室が公開している，「BERT の日本語事前学習モデル」[2]を用いた．この事前学習モデルに対して，用意した学習データを用いてファインチューニングを行うことにより，分類器を作成した．事前学習済み日本語 BERT モデル（tohoku-nlp/bert-base-japanese-whole-word-masking）を基にしたマルチラベル分類モデルを構築した．

5.2 実験結果・考察

小林らはモデルの精度指標として RMSE と MAE を用いているため，本稿でも同様の指標を用いる．また，推定制度を比較するために追加で正解率，適合率．再現率，F 値を算出した．分析結果を表 4 と表 5 に記す． 楽天トラベル（評価 7（総合）× ユーザ投稿本文）の結果は RMSE：0.41，MAE：0.17，楽天ゴルフ（総合評価 × コメント）の結果は RMSE：0.43，MAE：0.18，楽天市場（評価ポイント × レビュータイトル）の結果は RMSE：0.37，MAE：0.14，楽天市場（評価ポイント × レビュー内容）の結果は RMSE：0.36，MAE：0.13 となった．小林らの結果を上回る結果となったため，提案手法により類似したデータカラムを抽出することができ，そのデータカラムが分析技術に適用可能と示唆された．この検証では，データセットに対して，用いた技術が限られているため，多種多様な分析手法での検証が十分でない点が課題として挙げられる．そのため，他の分析技術での適用可能性について検討を行い，技術選択の汎用性を検証する必要がある．

本稿では，楽天データセットのみを使用しておりデータセットの内容に偏りがある懸念が存在する．そのため，今後は異業種のデータ（例：医療データや金融データ）への適用可能性を検証し，提案手法の汎用性を検証する必要がある．また，本実験において，BERT の学習データとして使用した評価値のデー

1：https://huggingface.co/tohoku-nlp/bert-base-japanese-v2
（2025/02/12 確認）

2：https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/
（2025/02/12 確認）

表 4: 分類タスクの結果

	楽天トラベル	楽天 GORA	楽天市場（レビュータイトル）	楽天市場（レビュー文）
正解率	0.53	0.34	0.58	0.57
適合率	0.38	0.35	0.36	0.43
再現率	0.25	0.19	0.25	0.31
F 値	0.26	0.24	0.25	0.34

表 5: 回帰タスクの結果

	楽天トラベル	楽天 GORA	楽天市場（レビュータイトル）	楽天市場（レビュー文）
RMSE	0.41	0.43	0.37	0.36
MAE	0.17	0.18	0.14	0.13

タは全てのデータセットで 5 段階評価のデータであったため，評価値推定の精度が高かった可能性がある．そのため，3 段階の評価値データや「良い・普通・悪い」などの順序尺度の当てはまるテキストデータでも同じような精度が得られるか検証する必要がある．

6 おわりに

本稿では，データの特性，適用する分析技術，および得られる結果の三者の関係性に注目し，既存の活用事例を分類整理した．提案手法として，類似した特性を持つデータには同じ分析技術が適用可能であるという仮説のもと，データセット内のカラムの形式と性質の類似性を基に適用可能な技術を推薦する方法を提案した．

提案手法の有効性を検証するために，楽天市場データセットを用いて評価ポイントの予測を実施した結果，過去の事例より高い精度を示した．これにより，本手法が一定の有効性を持つことが示唆された．また，メタデータによって推定された類似データカラムであれば比較元のデータカラムと代替できる可能性が示唆された．今後は，「データ × 技術 × 結果」の形式で整理したもののうち，一つが欠けていても他の二つの情報から推定するには，データカラムと同様に，用いる技術や目的である結果のそれぞれの類似度の算出方法を検討する．また，データ DB におけるメタデータを付与するプロセスは手動で行っているため，自然言語処理を用いた自動ラベリングなどの手法を活用し，メタデータ付与の自動化を進める具体的な方法を検討を行う．

謝　辞

本研究の遂行にあたり，国立情報学研究所の IDR データセット提供サービスにより楽天グループ株式会社から提供を受けた「楽天データセット」[3]を利用した．記して謝意を表す．

文　献

[1] 相原健郎: ビッグデータを用いた観光動態把握とその活用：動体データで訪日外客の動きをとらえる, 情報管理, Vol. 59, No. 11, pp. 743–754 (2017).
[2] 草刈祐子, 中藤哲也, 鈴木孝彦, 廣川佐千男: 過賞賛のレビューの検出と特徴語抽出, 第 17 回情報科学技術フォーラム, Vol. 17, No. F-016, pp. 255–256 (2018).
[3] 玄道俊, 松下光範, 山西良典: コンテンツと人の関係に着目したコンテンツデータセットの抽象化によるデータ利用傾向の俯瞰, No. E34-4 (2022).
[4] 小林義幸, 越仲孝文: EC サイトのレビューテキストからのレーティング予測と購買者評価の分析, Vol. JSAI2022, pp. 1P5GS602–1P5GS602 (2022).
[5] 酒井美春, 上田真由美, 松下光範: 化粧品の評価項目別スコア生成のための評価表現辞書の自動構築, 第 11 回データ工学と情報マネジメントに関するフォーラム, No. B6-2 (2019).
[6] 住田篤紀, 山田泰寛: レビュースコアとレビューの文字数に着目した商品スコアの再計算, 第 20 回情報科学技術フォーラム, Vol. 2, pp. 197–198 (2021).
[7] 高橋利知, 松浪佑宜, 柴田克彦, 川上理亮, 高原, 宮田翔平, 赤司泰義: 需給予測を用いたエネルギー自立型システムの構築: 長・短期記憶（LSTM）を用いた電力需要予測結果の評価およびケーススタディ, 技術報告 No.37, 高砂熱学イノベーションセンター (2023).
[8] 早矢仕晃章, 大澤幸生: Data Jacket Store: データ利活用知識構造化と検索システム, 人工知能学会論文誌, Vol. 31, No. 5, pp. A–G15_1–9 (2016).
[9] 樋口亮太, 山西良典, 松下光範: コミックの登場人物についての説明文からの性格タグ推定, ARG WEB インテリジェンスとインタラクション研究会予稿集, Vol. 16, pp. 112–115 (2020).
[10] 樋口亮太, 山西良典, 松下光範: 単語の頻度と意味に基づいたコミックに関するテキスト情報源の特性分析, 第 14 回データ工学と情報マネジメントに関するフォーラム, pp. E21–2 (2022).
[11] 大和大祐, 野村眞平: SUUMO でのビッグデータ活用事例, 日本不動産学会誌, Vol. 31, No. 1, pp. 78–83 (2017).
[12] 山西良典, 藤岡寛子, 西原陽子: 擬似コーパスを用いた飲食店レビューの観点の自動分類, 人工知能学会論文誌, Vol. 36, No. 1, pp. WI2–A_1–8 (2021).
[13] 山本京佳, 山西良典, 松下光範: 半教師あり NMF を用いた専門分野と講義の関係推定, 2021 年度人工知能学会全国大会（第 35 回）論文集, Vol. JSAI2021, No. 1I2-GS-4a-01 (2021).
[14] 楽天グループ株式会社: 楽天データセット. 国立情報学研究所情報学研究データリポジトリ. (データセット), DOI: https://doi.org/10.32130/idr.2.0 (2014).

3：https://rit.rakuten.com/data_release/