データセットの内容的特徴に基づく データ利活用の支援に関する研究

総合情報学研究科
知識情報学専攻

インタラクションデザインの理論と実践

23M7134
林 沙也加

論文要旨

企業活動で蓄積される大量のデータは，大規模データセットとして整理・活用されることで，ユーザの行動やニーズの推測に役立ち，ビジネスの貴重な情報源となる．データ分析には特性や構造を考慮し，適切な技術を選択することが不可欠だが，知識のないユーザは適切な手法を選定できず，既存の支援システムも新規データや多様な目的に対応しづらい．そこで，新規データにも柔軟に適用可能な技術推薦手法が求められる．本研究では，データ活用事例において，データ特性（e.g., レビューデータ），分析技術（e.g., BERT），得られる結果（e.g., 評価値推定）の三者関係に着目し,「データ × 技術 × 結果」の形式で整理・分類することで，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．この形式を採用することで，データ・技術・結果のうち二つの情報から欠けた要素を推定し，未知データの利活用を支援する．例えば，未知データセットが既存の活用事例と類似する場合，その「技術」と「結果」を推薦し，ユーザが適切な技術を選定できるようにする．これにより，レビュー文の特性から BERT の適用可能性を示し，技術に不慣れなユーザでもデータ特性に基づいた適切な技術選定が可能となる．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，カラムの形式と性質の類似性に基づくデータセット間の類似性を評価する．

データセットに形式や性質を表すラベルとしてメタデータを付与し，類似度を算出する手法を提案する．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，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 を実行するために必要なデータの特性として，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．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．上記の議論を踏まえ，各データカラムに設定するメタデータを形式に関する観点（e.g.,「テキスト」や「数値」），内容に関する観点（e.g.,「説明」や「定性」），の 2 つの観点で整理した．

提案手法により類似している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」の部分におけるモデルの精度を評価した．既存の活用事例である論文は，レビューテキストから BERT を用いてレーティング（星の数）を予測したものである．この論文では楽天市場データセットの「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おり，類似しているデータカラムとして，楽天トラベルデータセットの「ユーザ評価の評価 7（総合）」や「ユーザ投稿本文」などが抽出できた．対象とした論文では，評価指標として二乗平均平方根誤差（RMSE）と平均絶対値誤差（MAE）を用いていた．そのため，本研究でもこれに倣い同様の評価指標を用いた結果，既存の活用事例の結果と遜色ない結果となったため，メタデータにより推定した類似データカラムであれば，代替可能である可能性が示唆された．

目 次

1 序論 1
1.1 データ活用の現状と課題 1
1.2 本研究が解くべき課題 3
1.3 本研究の目的 4

2 関連研究 7
2.1 データの利活用知識の構造化に関する研究 7
2.2 データセットの抽象化に関する研究 7
2.3 本研究の立ち位置 8

3 データセットの類似度 9
3.1 「データ × 技術 × 結果」の形式に整理する手法 9
3.2 データセットの類似度推定手法 9
3.3 実装手順 13

4 実験例 1：[BERT] × [数値：評価] と [テキスト：評価] の事例 17
4.1 活用事例の概要（小林ら） 17
4.2 実験手順 17
4.3 実験結果・考察 18

5 実験例 2：[k-means++法] × [文章]，[カテゴリ]，[日付] の事例 20
5.1 活用事例の概要（小川ら） 20
5.2 実験手順 20
5.3 実験結果・考察 22

6 議論 28
6.1 メタデータによる類似カラムの推定 28
6.2 メタデータの妥当性 29
6.3 データ・技術・結果の三要素による体系的整理の効果 31

7 結論 32

A 付録 34

1　序論

本章では，本研究の実施に至った経緯を述べ，対象とする課題を明確化する．

1.1　データ活用の現状と課題

1.1.1　企業のデータ利活用について

企業活動において蓄積される大規模データセットである商品目録データや販売データの利活用は，サービスの開発や提供の実現に不可欠であり，労働や資本と同様に重要な要素となっている [3]．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である．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することができる．

大規模データセットから様々な有益な情報を取得することができるが，企業がデータを活用しきれていないという課題が存在する [21]．総務省が行った各国企業の「パーソナルデータ（e.g.， 顧客の基本情報）」の活用状況の調査結果（図 1.1）によると，「活用できている」（「既に積極的に活用している」と「ある程度活用している」の合計）と回答した日本企業の割合は 52.8%であり，諸外国の企業と比較すると割合が低い．企業がデータを活用できていない理由として，日本企業は「データの利活用方法の欠如，費用対効果が不明瞭」，「データを取り扱う（処理・分析等）人材の不足」と答えた割合が諸外国と比較すると多かった（図 1.2）．このように，企業内で十分な研究開発力を備えているところが多いとはいえない状況である．

1.1.2　研究機関のデータ利活用について

ビッグデータ処理技術やディープラーニングなどの技術的な進歩に伴い，情報学研究においても大規模データセットの活用が不可欠となっている．また，実用的な研究成果を残すためにも研究機関には実社会で生成された大規模なデータセットが必要である．しかし，

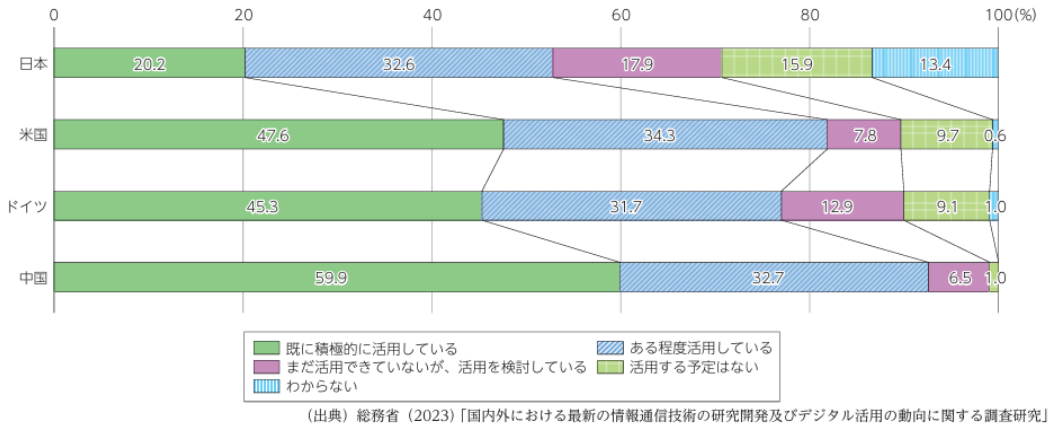

図 1.1: 各国企業におけるパーソナルデータの活用状況（文献 [21] より図引用）

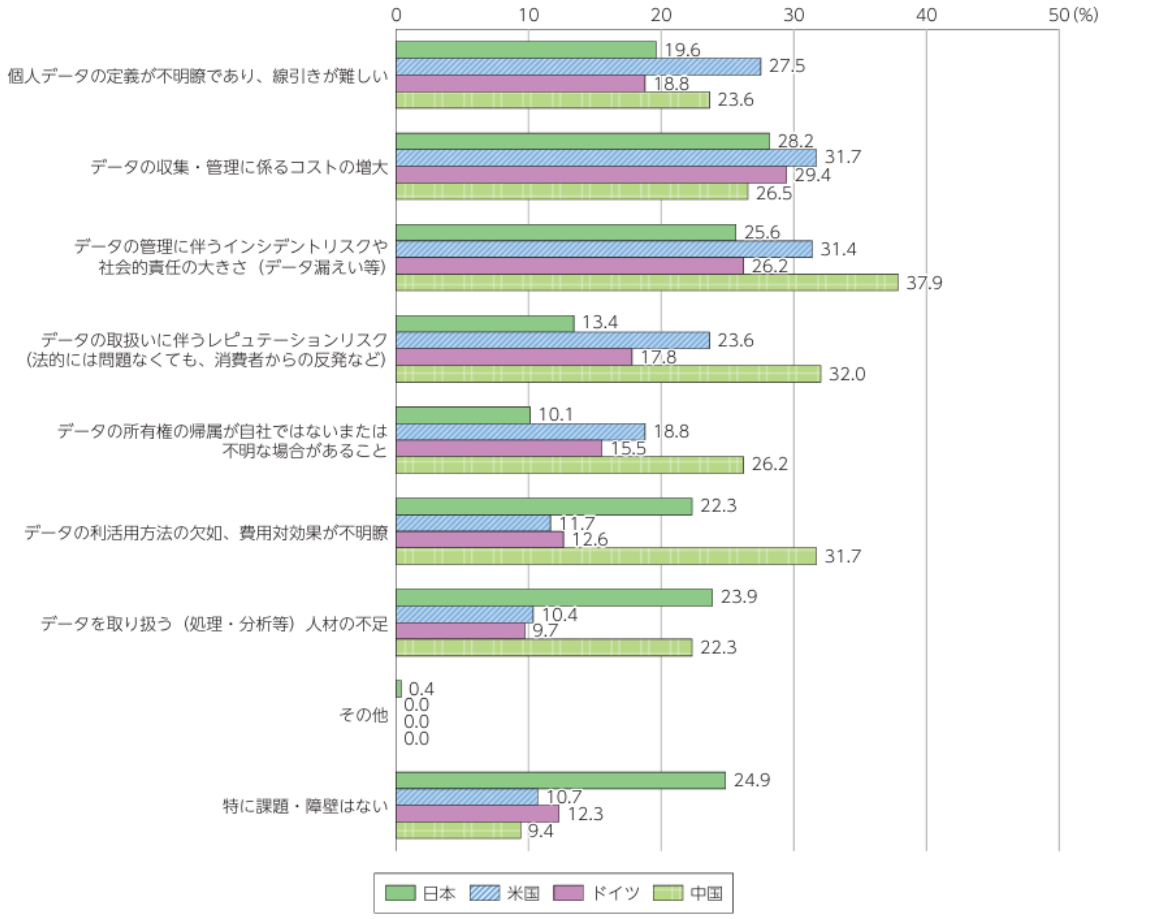

図 1.2: パーソナルデータの取扱いや利活用において想定される課題や障壁（文献 [21] より図引用）

大学や研究機関では十分なデータを確保することが困難であるという課題がある．

データの利活用が可能な技術を有する研究者が研究を一から始める際，研究者自身で目的に応じたデータを収集し，データセットを構築する．しかし，データセットを構築する上で，以下のような問題点が指摘されている [14]．

- 十分な規模を確保するためにコストがかかる
- 内容や形式を偏りなくデータを集めることが難しい
- データセットの構築方法や使用方法の適法性が明らかでない場合がある
- データに含まれる未処理の著作権や個人情報などが存在する可能性がある

これらを解決するために，民間企業のような法的責任や技術力を持つ機関と協力して，企業が構築したデータセットを研究者に提供するといった解決方法がある．しかし，この問題点として協力関係にいない他の研究者がデータセットを使用できず研究成果を検証できない点やデータセットに統一性がないことで相互比較ができない点が挙げられる．

1.1.3 データの共同利用

1.1.1 節と 1.1.2 節で述べたように，企業と研究者のデータセットの共同利用を円滑に進めるための仕組みが求められている．データセットの共同利用を目的として，公共機関，民間企業，学術機関，市民団体などが社会問題の解決のためにデータを共有・活用する新しいコラボレーションモデルであるデータ・コラボラティブ [6] やデータの価格やデータ共有者の許可なく自由に再利用できるようにするオープンデータ [8] などが挙げられる．国立情報学研究所 (NII) データセット共同利用研究開発センター (DSC) では，産学界から受け入れた研究用のデータセットを研究者へ提供する窓口として，情報学研究データリポジトリ (IDR) を運営しており，2020 年 8 月の時点で 12 企業 26 種類のデータセットを提供している [12, 13]．また，このようなデータセットを用いた研究 [4, 9, 10] も増加している．

企業や研究者がデータを活用する機会が増えているが，データの活用には依然として課題が残されている．例えば，データの形式や内容が多様化する中で，ユーザは自身の目的に適したデータを選定し，データに適用可能な分析手法を見極めることが求められる．しかし，すべてのユーザがデータ処理や分析技術に精通しておらず，特に新規のデータセットを活用する際には適切な分析技術の選定が困難になることがある．そのため，データ特性（e.g.，評価や感想などの内容特性，数値やテキストなどの形式特性）を考慮した適切な技術選定を支援する枠組みが必要である．

1.2 本研究が解くべき課題

1.1 節で述べたように，大規模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 [27] 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 [11] などが挙げられる．データセットをユーザの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．

こうした課題を解決するためには，過去の活用事例を参照し，類似した方法を見つけて適用することが有効である．しかし，データ活用の目的や適用可能な技術は数多く存在するため，ユーザの目的に沿った適切なデータや技術を見つけることは容易ではない．そこで，本研究では，既存の活用事例のデータと新規のデータが形式的・内容的に可換である場合，適用可能な技術を特定できる，という仮説に基づき，活用事例と新規データの対応関係を分析し，適切なデータ活用技術を推薦する手法を提案する．本手法により，データの特性を考慮した適切な分析技術の選定を支援し，新規のデータセットにも柔軟に適用可

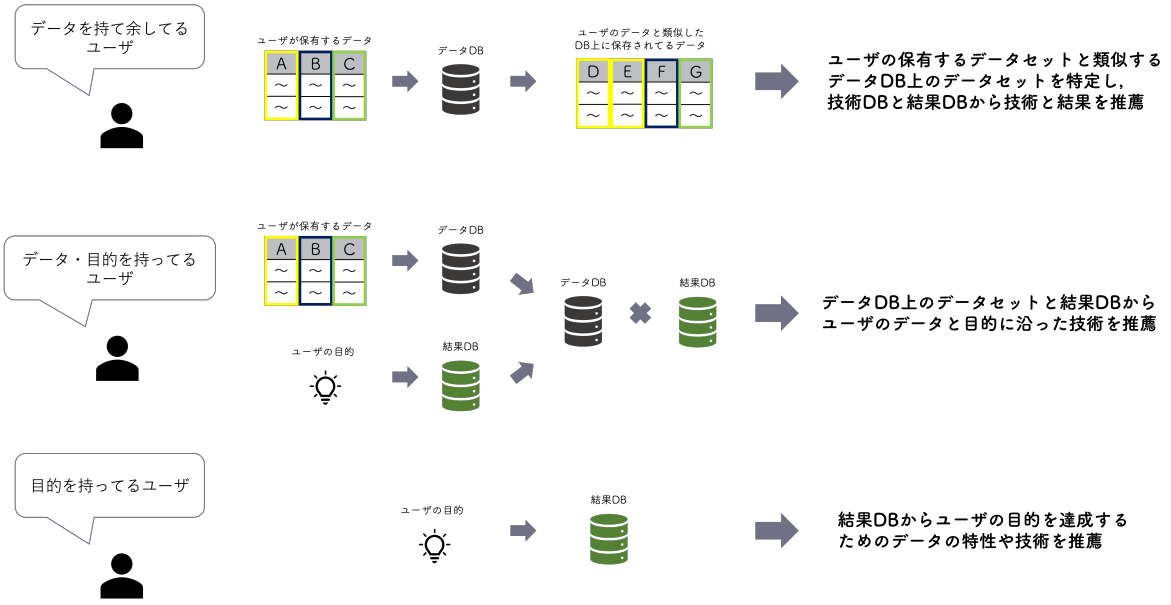

図 1.3: 活用例の図

能なデータ活用の促進を目指す．

1.3 本研究の目的

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，既存の活用事例を「データ × 技術 × 結果」の形式で整理・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．上記した得られる結果とは，ユーザの保有するデータからユーザが期待する漠然とした結果（e.g.， 評価値を予測したい，カテゴリごとのレビューの特徴を知りたい）のことを指す．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，ユーザが新規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．

ユーザが,「データ × 技術 × 結果」の形式で整理された仕組みを活用する例を図 1.3 に示す．活用例は，以下のように説明することができる．

新規のデータセットの場合

ユーザが持つ新規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（新規）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新規のものに細分化される．

- 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新規の場合，新規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（新規）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．

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．新規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，新規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．

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 [18] 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

表 1.1: 活用事例を「データ × 技術 × 結果」の形式でまとめた例

論文名	データ	技術	結果
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 [18]	評価値 レビュー	BERT リッジ回帰 RNN	評価値の予測
半教師あり NMF を用いた 専門分野と講義の関係推定 [29]	卒業論文概要 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
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 [20]	商品レビュー	文字数のカウント	評価値の予測
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 [19]	化粧品レビュー 評価値	Word2Vec	評価表現辞書の構築
過賞賛のレビューの検出と特徴語抽出 [16]	評価値 レビュー	SVM	過賞賛レビューの検出
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 [25]	キャラクタ説明文 性格タグ	ニューラルネットワーク	性格タグを推定
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 [26]	キャラクタ説明文 レビュー文	クラスタリング	各特徴の明確化
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 [28]	クックパッドのつくれぽ 楽天トラベルのレビュー	fastText	レビュー観点の分類
商品レビューにおける時系列情報に着目した クラスタ分析と可視化 [15]	レビュー 属性情報	k-means++法	属性ごとのレビューの特徴

事例 [22] 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 1.1 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．その一端として，本研究では，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目的で，データセットの研究利用に関する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確にする．

2.1　データの利活用知識の構造化に関する研究

データにメタデータをつけることでデータの概要情報を記述する研究が行われている．Ohsawa らは，データそのものではなく，変数名，保存形式，収集方法などの概要情報を記述フレームワークとしてデータジャケットを提案している [7, 23]．データジャケットは人が読むことを前提として自然言語で記述されており，ユーザはデータの中身に直接アクセスすることなく，その利用価値を検討することが可能となる．個人情報の存在により共有が困難なデータに対してデータジャケットを用いることで，セキュリティ上のリスクを低減させて，情報共有が可能となる．また，データジャケットは構造化されているため，人間だけでなく計算機にも可読性を有し，データの管理や活用の効率化が可能となる．

早矢仕らは，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 [24]．さらに，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．

2.2　データセットの抽象化に関する研究

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 [17]．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，それに基づいて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．この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ため，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，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（e.g.， 離散値–テキスト形式），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を横断的に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を十分に考慮していない点も指摘されている．

2.3 本研究の立ち位置

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．また，データセットの類似度を推定する手法を提案する．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応することで，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　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．

3.1　「データ × 技術 × 結果」の形式に整理する手法

1.3 章の図 1.3 について,「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 3.1 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．

（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 3.1-④のデータ DB，技術 DB，結果 DB から，図 3.1-①，②，③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．

3.2　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 [28]．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，

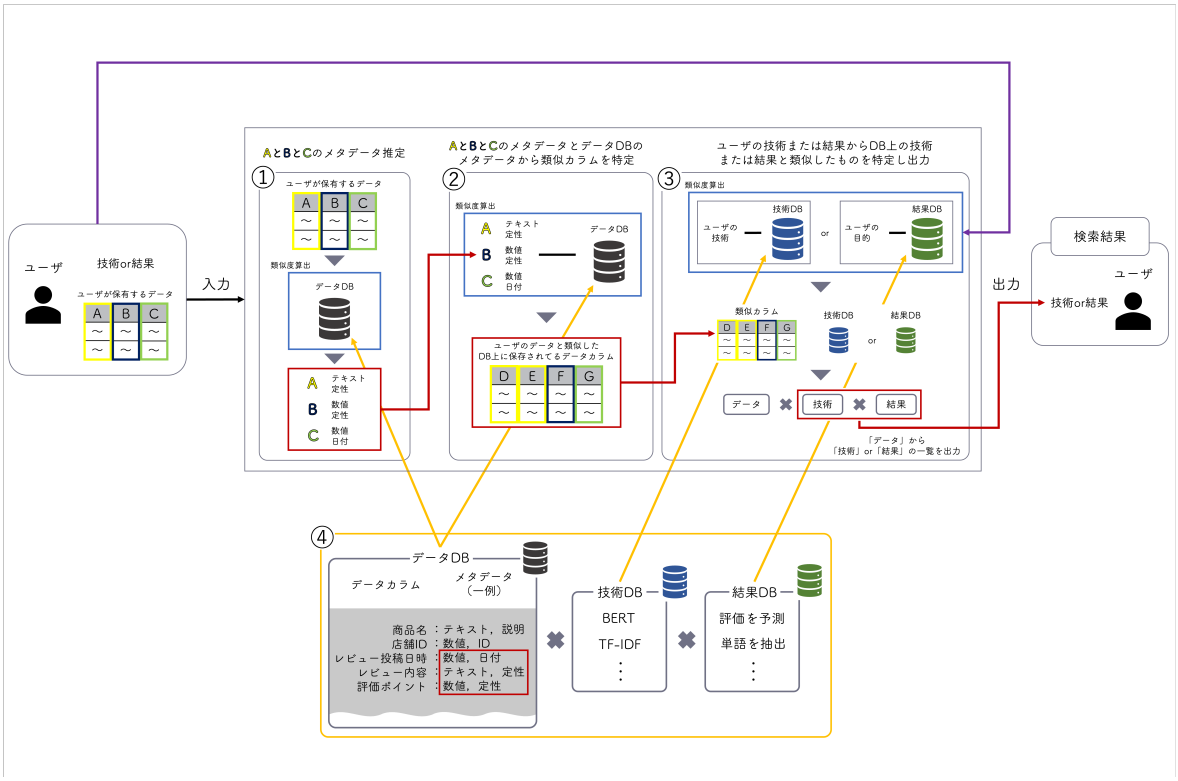

図 3.1: ユーザの新規データ利活用の図

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．

3.2.1 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，評価値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．これらのデータはテキストや数値という特徴だけでなく，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することができるため，これを考慮した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 3.2，図 3.3 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

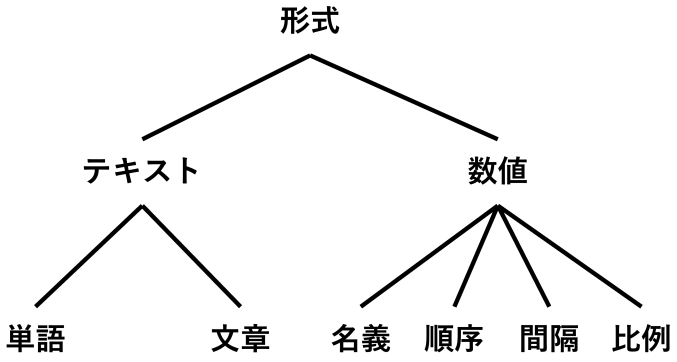

図 3.2: メタデータの構造（形式）

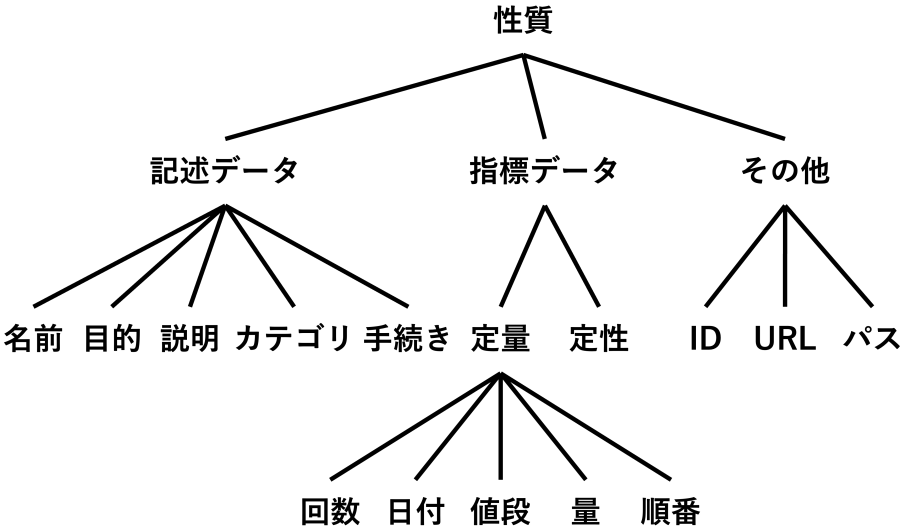

図 3.3: メタデータの構造（性質）

  - **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）

3.3 実装手順

本稿では，3.2 節で定義したメタデータを楽天データセット [30] 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，これらは研究者に向けて公開されており誰でも容易に入手できる点や，各データセットは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利用されており，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 3.1 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．

（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「記述データ」（e.g., 名前，説明，目的）や「指標データ」（e.g., 定性，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．

表 3.1: 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

(a) 楽天市場データセット

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

(b) 楽天レシピデータセット

レシピ情報	レシピ ID，ユーザ ID，大カテゴリ，中カテゴリ， 小カテゴリ，レシピタイトル，レシピのきっかけ， レシピの紹介，料理画像ファイル名，料理名， タグ 1，タグ 2，タグ 3，タグ 4， ワンポイント情報，調理時間の目安 ID， どんな時用 ID，費用の目安 ID，何人分， レシピ公開日
材料情報	レシピ ID，素材名，分量
作り方情報	レシピ ID，手順位置，手順作り方
「つくったよ」レポート情報	レシピ ID，ユーザ ID，おすすめコメント， オーナーコメント，作成日時
Pickup レシピ	ピックアップ日，レシピ ID
デイリシャスニュース	ニュース記事 ID，公開日時，ライター ID， ジャンル名，記事タイトル， 記事テキスト（メイン記事）

(c) 楽天トラベルデータセット

ユーザ評価	投稿者 ID，投稿日時，施設 ID，プラン ID， プランタイトル，部屋種類，部屋名前，目的， 同伴者，評価 1（立地），評価 2（部屋）， 評価 3（食事），評価 4（風呂），評価 5（サービス）， 評価 6（設備），評価 7（総合），ユーザ投稿本文， 施設回答本文
施設名マスタ	施設 ID，施設名

(d) 楽天 GORA データセット

クチコミ情報	クチコミ ID，コース ID，クチコミ投稿者名， 都道府県，年齢，平均スコア，オススメ目的， オススメタイプ，利用回数，総合評価， コストパフォーマンス，スタッフ接客， コース/戦略性，食事が美味しい，設備が充実， フェアウェイが広い，距離が長い，タイトル， コメント，クチコミ登録日，プレー日

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「指標データ」と「定性」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 3.2，表 3.3，表 3.4，表 3.5，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．なお，他のデータセットカラムの結果については付録 A に示す.

表 3.2: 「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

表 3.3: 「商品データ」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
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
店舗コード	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
商品コード	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
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
商品価格	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	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
商品画像 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
販売方法別説明文	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
商品説明文	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

表 3.4: 「ジャンルマスタ」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
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
親ジャンル 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
ジャンル名	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

表 3.5: 「市場店舗レビューデータ」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
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
レビュー本文	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
レビュー投稿日時	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

4　実験例1：[BERT] × [数値：評価] と [テキスト：評価] の事例

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を用いて，既存の活用事例を「データ，技術，結果」に整理したもののうち,「データ」の部分に類似データを代用した際のモデルの精度を評価する．本研究では各分析技術の妥当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．本章では，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．

4.1　活用事例の概要（小林ら）

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 [18]．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」を使用しており，商品カテゴリを「白米」に限定している．また，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）[2] を採用し，それぞれの予測精度を比較した．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]を用いた．また，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．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．

4.2　実験手順

小林らが行った研究の構造に基づいた再現するために BERT を用いて類似データの代用可能性を検証した．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する．評価指標として，RMSE と MAE を本研究でも用いる．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 以外にもリカレントニューラルネットワーク（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．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 値の値も算出する．

[1] https://huggingface.co/tohoku-nlp/bert-base-japanese-v2（2025/02/14 確認）

小林らの研究を再現するために，3.2 章で述べたデータカラムの推定手法を用いて推定した類似するカラムを抽出した．表 3.2 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 3.2 中の 4，17 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 3.2 中の 6，17 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推測した.「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推定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

データカラムに含まれるデータをそれぞれ 10,000 件ずつ用意し，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）をもとにしたマルチラベル分類モデルを構築した．

4.3 実験結果・考察

小林らの結果を参考に，RMSE と MAE の結果と推定制度を比較するために追加で機械学習モデルの性能指標である正解率，適合率，再現率，F 値を算出した．正解率とは，全ての予測の中で正しく分類された予測数の割合を示す指標で，データの偏りに影響を受けやすくいというデメリットがある．適合率とは，モデルが「正しい」と判断した予測の答えが「正しい」となった割合を表す指標で，予測のミスを減らしたい場合に適している．再現率とは，答えが「正しい」であるもののうち，モデルが「正しい」と判断した割合を表す指標で，正しいデータを漏れなく見つけたい場合に適している．適合率と再現率は，片方が高くなるともう片方が低くなる．二つの値のバランスをとる指標として F 値があり，適合率と再現率の調和平均をとることで計算される．これらの値を算出することで，モデルの精度を検証する．

分析結果を表 4.1 と表 4.2 に記す． 表 4.2 より，小林らの結果と遜色のない結果となった．そのため，提案手法により推定された類似データであれば，既存の活用事例で使用されたデータと代替可能である可能性が高く，それに適用された分析技術に類似データが適用可能であると言える．

[2] https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/（2025/02/14 確認）

表 4.1: 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 4.2: 回帰タスクの結果

	小林らの結果	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	0.60	0.41	0.43	0.37	0.36
MAE	0.36	0.17	0.18	0.14	0.13

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見られると考えられる．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高める必要がある．今回 BERT で使用した評価値のデー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良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．

5　実験例2：[k-means++法] × [文章]，[カテゴリ]，[日付] の事例

本章では，小川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．この事例では，過去事例と

5.1　活用事例の概要（小川ら）

小川らは，商品レビューにおける時系列情報と属性情報に着目し，クラスタリング手法を用いてレビューを分類することでレビュー全体の傾向を分析した [15]．使用しているデータカラムは，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「性別」，「レビュー登録日時」,「レビュー内容」であり，商品カテゴリを「デジタルカメラ」に限定した．レビューの特徴量として，TF-IDF とベクトルが高次元にならないよう次元削除手法の一つである LSA(Latent semantic analysis) によって次元数を削除した文書ベクトルを用いた．また，各レビューの中で TF-IDF の値が最も高い単語をそのレビューの最重要単語とし，word2vec[5] を用いて算出した最重要単語の単語ベクトルを特徴量として使用した．さらに，時系列情報ベクトル，属性情報ベクトルを抽出し，k-means++法 [1] によるクラスタリングを行った．ここで，クラスタリングの距離尺度にはコサイン類似度を使用した．可視化には SOM を用いてクラスタ間の関係性を保持した可視化を行ったほか，条件付き確率の比を指標とすることで特徴的なクラスタを抽出した．結果を表 5.1，表 5.2 に記す．実験では，デジタルカメラのレビューを対象に分析を行い，例えば 2012 年 3 月のレビューでは，男性に「新機種」や「最新」，女性に「卒業式」や「美肌」などの特徴的なキーワードが現れることが確認された．SOM による可視化はクラスタの関係性を把握するのに有効であり，word2vec によるキーワード抽出は意味的に類似した単語の分類に有効だとわかった．

5.2　実験手順

小川らが行った研究 [15] に基づいた再現によって k-means++法を用いて類似データの代用可能性を検証した．小川らの研究では，楽天市場データセットの「みんなのレビュー・口コミ情報」に含まれる「性別」データカラムを用いてレビューを「男性」と「女性」に分類し，それぞれの単語の出現傾向を分析した．楽天市場データセットには，かつては「性別」データカラムが含まれており，小川らの研究ではそれを利用していた．しかし，本研究で用いる楽天データセットの各種データセット（e.g., 楽天トラベルデータセット，楽天レシピデータセット）には,「性別」データと同様のデータを含むデータカラムは存在しておらず，メタデータが完全一致するものは存在しなかった．この「性別」データカラムは「男性」と「女性」の二つのデータが含まれており，このデータカラムを基準とすることで,「性別」データカラムが含まれるデータセット中のデータを「男性」と「女性」にカテゴリ化し分割できる．小川らの手法では，この性質を踏まえ，男性か女性かでレビュー文中の出現単語の傾向が異なるか，を見るために用いている．そのため，データをカテゴリごとに分類する際に基準となりえるデータ（以下，カテゴリカルデータと記す）であれば，この「性別」データカラムと同等の役割を果たすと考えられる．そこで，本研究では「性別」データの

表 5.1: 男性に特徴的なクラスタ群の一例（文献 [15] より表引用）

期間	クラスタのキーワード
3 月	新機種，機種，被写体，ドット，コンデジ，カセット，録音，今頃，工事写真，ライティング， シンプル，デザイン，携帯，画質，格好，機能，カッコイイ，マッチ，レトロ，カッコ， 最新，後継，新型，進歩，搭載，旧型，一昔，発売，モデルチェンジ，進化
10 月	ゴルフ，バイク，アウトドア，散歩，競技，ランニング，体育館，観戦，行楽，登山， 電池，乾電池．単三．予備．寿命．充電．バッテリ．充電池．消耗．機種， マクロ，接写，夜景，野鳥，撮影，顕微鏡，手ブレ，レンズ，機能，パノラマ
12 月	時間，経過，以内，充電，撮影，半日，開始，一週間，連続，週間， 録画，起動，動作，削除，再生，メモリ，ソフトウェア，認識，パソコン，画像， 倍率，ピント，アングル，望遠，無限遠，不満，カメラ，レンズ，必要，コンパクトデジカメ

表 5.2: 女性に特徴的なクラスタ群の一例（文献 [15] より表引用）

期間	クラスタのキーワード
3 月	美肌，メイク，成果，オンチ，小顔，効果，化粧，クマ，シミ，把握， 卒業式，入園，入学式，入学，子供，卒園，写真，小学校，カメラ，七五三， ピンク，ホワイト，ピンク色，黄色，ブラック，レッド，ブラウン，茶色，タッチパネル，ビビッド
10 月	運動会，行事，花火大会，幼稚園，子供，遠足，音楽会，花火，保育園，一眼レフ， 荷物，収納，整理，旅行，一眼レフ，タクチン，パンパン，レジ袋，母親，容量， 簡単，便利，調節，片手，スライド，機能，可能，操作，最初，ワンタッチ
12 月	クリスマスプレゼント，息子，デジカメ，ピンク，クリスマス，愛用，値段，子供，電池，年齢， プリント，写真，注文，大量，用紙，アルバム，自宅，インク，サイズ，出方， ペット，黒猫，愛犬，愛猫，ワンコ，動物，子猫，子犬，ネコ，撮影

代わりに他のカテゴリカルデータを用いることで，同様の分析が可能か検討する。たとえば,「10 代，20 代，30 代，40 代」のようなデータが含まれる「年代」データカラムが商品レビューデータセット中にあるとすれば，10 代であれば「勉強，学校，流行」などの単語が多く出現し，20 代であれば「旅行，仕事，プレゼント」などの単語が多く出現する，といったレビュー中の出現単語の傾向が異なるかを見ることができる．このように，カテゴリカルデータであれば,「性別」データカラムを代替することができると考えた．楽天データセットの中でカテゴリカルデータとして使用できる可能性のあるデータカラムとして，楽天トラベルの「同伴者」,「目的」と楽天市場の「頻度」,「使い道」,「目的」が代替可能ではないかと推測した．これらのデータカラムに付与されたメタデータの形式的特徴として「テキスト（単語)」，性質的特徴として「記述データ（目的)」と「指標データ（定量（回数))」が付与されている．以下にこれらのデータカラムのデータ内容を記す．

- 付与されたメタデータが「テキスト（単語)」と「記述データ（目的)」
  - 楽天トラベル :「同伴者」(「一人，家族，仕事仲間，友達，恋人」)
  - 楽天トラベル :「目的」(「レジャー，ビジネス，その他」)
  - 楽天市場 :「使い道」(「イベント，おもたせ・ギフト，ビジネス，プレゼント，実用品・普段使い，趣味」)

  - 楽天市場 :「目的」(「家族へ，仕事関係へ，子供へ，自分用，女性（彼女、妻）へ，親戚へ，男性（彼、夫）へ，友人へ」)

- 付与されたメタデータが「テキスト（単語)」と「指標データ（定量（回数))」
  - 楽天市場 :「頻度」(「はじめて，リピート」)

「性別」データが存在していた場合，付与されるメタデータは，形式的特徴として「テキスト（単語)」，性質的特徴として「記述データ（カテゴリ)」であると考えられる．このように，形式的特徴を表すメタデータは一致していたが，性質的特徴を表すメタデータは一致しなかった．そのため，現在は完全一致したメタデータを持つカラム同士を類似カラムとしているが，その一致条件を緩和することで代替可能なデータカラムを推定することが可能であると考えた．

また，本稿では小川らの手法の中でも「性別」,「レビュー内容」の組み合わせを採用し，実験を行った.「レビュー内容」と類似するものとして，以下のカラムを使用した．

- 楽天市場データセットの「レビュータイトル」
- 楽天トラベルデータセットの「ユーザ投稿本文」

小川らはデータカラムの商品カテゴリを「デジタルカメラ」に限定しているが，本研究では楽天市場と同じ内容のデータのみを扱うわけではないため,「デジタルカメラ」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．本研究では，上記の代替可能データカラムを使用し，それらがレビュー中の単語の出現傾向に影響を与えるかを分析する．

5.3 実験結果・考察

実験結果を表 5.3，表 5.4，表 5.5，表 5.6，表 5.7 に記す．楽天トラベルの「同伴者」カラムのレビュー文中の単語の特徴として,「一人」のカテゴリでは「シングル，カプセル」など，一人で宿泊するユーザのレビューを代表する単語が現れている.「仕事仲間」のカテゴリでは「同僚，上司，サラリーマン」など，仕事関連の関係性を示す単語や,「会社，出張，仕事」といったビジネス用途に関連する単語が見られる.「友達」のカテゴリでは「友人，友達」といった関係性を示す単語に加え,「飲み会，ツイン，セミ，おしゃれ，ディズニー」といった，友人同士の宿泊やレジャーを表した単語が含まれている.「家族」のカテゴリでは「子連れ，息子，家族，子供，親子，添い寝」など家族構成を示す単語や,「布団，ベッド，ツイン，ダブル，セミ」といった人数が多いユーザを表す単語が見られる．また,「休日，クリスマス，誕生，usj，ディズニー」といった家族向けイベントやテーマパーク関連の単語も含まれている.「恋人」 のカテゴリでは「高層，イベント，シャトル，リゾート，おしゃれ」といった恋人と宿泊する際に着目する宿泊施設の特徴を表す単語や,「クリスマス，誕生」といった記念日関連の単語が目立つ．

楽天トラベルの「目的」カラムのレビュー文中の単語の特徴として,「ビジネス」のカテゴリでは「予定，仕事，次回，ビジネス，設備，金額，滞在，疲れ」など，出張や仕事に関する単語や「会社，同僚，常宿」など仕事関係の宿泊者を想定した単語が多く含まれてい

表 5.3: 楽天トラベルの同伴者に特徴的なクラスタ群の一例

同伴者	クラスタのキーワード
一人	シングル，カプセル，ベッド，立地，条件，朝食，内容，寝心地
仕事仲間	同僚，上司，サラリーマン，リラクゼーション，会社，出張，仕事
友達	友人，友達，飲み会，ツイン，セミ，おしゃれ，ディズニー
家族	子連れ，息子，家族，子供，布団，親子，添い寝，休日，ツイン， ダブル，ベッド，セミ，クリスマス，誕生，usj，ディズニー
恋人	高層，イベント，シャトル，リゾート，おしゃれ， クリスマス，誕生，バー，ベッド，セミ，ツイン，ダブル

表 5.4: 楽天トラベルの目的に特徴的なクラスタ群の一例

目的	クラスタのキーワード
ビジネス	予定，仕事，次回，ビジネス，設備，金額，滞在，疲れ，福岡， 広島，会社，寝心地，近隣，長期，常宿，飲み屋，新宿，遅く， 同僚，宇都宮，宮崎，朝飯，破格，長崎，水戸，山口，立地
レジャー	ゴルフ，usj，奈良，高松，子連れ，リフレッシュ，エステ， 那覇，感謝，景色，入浴，通り，ディズニー，成田，２人，両親， 神戸，パーク，レストラン，ベット，時期，暖房，館内，和室，布団
その他	アメニティ，個人，臭い，空調，タクシー，親戚，知人， トリプル，ファミマ，冷房，受験，空室，飲み会，ステイ， 忘年，都内，洋室，改装，専用，風呂場，来年，室料

る．さらに，「福岡，広島，宇都宮，宮崎，長崎，水戸，山口，新宿」など国内の主要都市の地名が多く登場しており，ビジネス利用者の宿泊先の分布を反映していると考えられる．「レジャー」のカテゴリでは「ゴルフ，usj，ディズニー，パーク，レストラン」など，観光や娯楽を目的とした宿泊を示す単語や「奈良，高松，神戸，沖縄（那覇）」などの観光地の地名も見られる．「子連れ，両親，和室，布団」といった単語からは，家族旅行向けの宿泊施設の特徴を表す単語や，「リフレッシュ，エステ，景色，入浴，館内」といったリラックスや癒しについて表す単語も多いことがわかる．「その他」のカテゴリでは「アメニティ，空調，タクシー，風呂場，室料」など，宿泊施設の設備やサービスに関する単語が見られる．また，「個人，親戚，知人」などの個人的な理由での宿泊を示す単語や，「受験，専用」など特定の目的での利用を示す単語も含まれている．

楽天市場の「頻度」カラムのレビュー文中の単語の特徴として，「はじめて」のカテゴリでは「品揃え，繁忙，セール，正規，純正」など，初めて購入・利用する際の品揃えや販売状況に関連する単語が多く見られる．また，「彼氏，友達，旦那，家内，姪っ子，甥っ子，主人，父親，祖母」など，特定の家族や知人への贈り物として利用する特徴があることがわかる．「誕生，クリスマス，年始，歳暮，帰省，土産，おせち」など，特別なイベントやギフト需要に関連する単語も多く登場している．さらに，「肌触り，履き心地，締め付け，着心

表 5.5: 楽天市場の頻度に特徴的なクラスタ群の一例

頻度	クラスタのキーワード
はじめて	品揃え，繁忙，彼氏，通学，マッサージ，毛穴，先輩，友達，歳暮， 旦那，配達，誕生，セール，肌触り，履き心地，年始，甥っ子， 相手，家内，正規，帰省，土産，締め付け，姪っ子，主人，心配， 翌日，コンパクト，純正，フィット，音質，首回り，祖母，褒美， リクエスト，値打ち，着心地，手触り，通勤，大量，オシャレ， おせち，写真，期待，機能，希望，収納，クリスマス，父親
リピート	世話，両親，シリーズ，会社，昨年，在庫，肩幅，付け心地， 得用，指通り，毎月，欠点，備品，毎回，定期，地元， 文句，年賀，毛玉，定番，継続，送付，ストック，薄型，半端， 常備，知り合い，吸水，毎年，仕事，連絡，コスパ，ポイント， 再度，リピーター，まとめ買い，リピ買い，枚数，受注，他店， 以前，年末，便利，普段，最高，今年，来年，前回，クーポン

地，手触り，音質，フィット」など，初めて使用する商品の品質や快適さを意識する傾向が見られる.「期待，値打ち，希望，心配」など，購入前後の期待感や不安に関する単語も含まれている.「リピート」のカテゴリでは「シリーズ，定番，継続，常備，ストック，定期，まとめ買い，世話」など，継続して利用する商品やサービスに関連する単語が多く見られる．また,「毎回，毎年，毎月，年末，昨年，今年，来年，前回，リピーター，リピ買い，以前」など，定期的な購入や利用に関する単語が多く登場している.「会社，仕事，備品」など，個人だけでなく仕事で使用する日用品などを意識したリピート利用も見られる．また,「文句，欠点，普段，便利，最高」など，リピート購入の際の評価や使用感を示す単語が含まれており，購入者の満足度や使い勝手に関するユーザの特徴が見られる．

楽天市場の「使い道」カラムのレビュー文中の単語の特徴として,「おもたせ・ギフト」 のカテゴリでは「贈答，年賀，熨斗，先方，品数，高価」など，贈り物や手土産としての利用を示す単語が多く見られる．また,「祖父母，兄弟，親族，義父，義母，おば，祖父」など，家族や親戚向けのギフトが多いことや,「カステラ，プリン，ハム，チョコレート，干し柿，フルーツ，ビール，菓子」など食品系のギフトが多いことがわかる.「来客，集まり，退職，田舎，紙袋」など，人が集まる場面での利用を想定した単語も含まれている.「イベント」のカテゴリでは,「新年，初詣，祝い，御節，忘年，パーティー」など，年末年始やお祝いごとに関連する単語が多く見られる．また,「ケーキ，寿司，焼き肉，しゃぶしゃぶ，毛蟹」など，イベント用の食事に関する単語も多いことがわかる.「ビジネス」 のカテゴリでは,「必要，業務，仕事，出張，管理，現場，書類，スーツ，バッグ，フォルム，合皮，上着，試着」など，仕事に関連する単語やビジネス用のファッションやアイテムに関する単語も見られる．また,「郵送，刻印，他社，判断」など，業務上のやりとりや品質の評価に関連する単語が多いことも特徴的である.「プレゼント」 のカテゴリでは,「彼氏，姪っ子，甥っ子，家内，知人」など，特定の人に向けたプレゼントとして利用されると思われる単語や.「誕生，還暦，出産，通学，受験，景品」など，特別なイベントに関連する単語が多く見られる．

表 5.6: 楽天市場の使い道に特徴的なクラスタ群の一例

使い道	クラスタのキーワード
おもたせ・ギフト	贈答，年賀，夫婦，ビール，集まり，退職，干し柿，おば， 祖父母，プリン，カステラ，知り合い，縁起，来客，紙袋， 兄弟，田舎，ハム，親族，義父，日頃，義母，先方，菓子， 熨斗，チョコレート，評判，高価，品数，フルーツ，祖父，
イベント	ケーキ，新年，祝い，初詣，毛蟹，焼き肉，パーティー， 御節，旅行，しゃぶしゃぶ，イベント，忘年，和洋，寿司， キャラクター，合格，手作り，ツリー，スニーカー
ビジネス	必要，首周り，ｐｃ，フォルム，合皮，郵送，バックル， 判断，高級，業務，出張，書類，スーツ，伝票，刻印， 上着，仕事，バッグ，安物，試着，管理，現場，他社
プレゼント	彼氏，受験，通学，冷え性，実印，還暦，予算，ベルト， ブランド，小学，誕生，実家，景品，甥っ子，出産，家内， 姪っ子，知人，ゴルフ，こども，大人，好み，気持ち，事前
実用品・普段使い	はき心地，付け心地，指通り，毎月，布団，シャツ， タオル，靴下，シャンプー，愛用，試し，まとめ買い， 品質，財布，音質，継続，リピーター，持ち運び， 必需，ストック，保湿，着心地，普段，使い心地， 毛布，美容，組み立て，締め付け，予備，重宝，弁当
趣味	中古，軽量，長持ち，プレイ，ディスク，レンズ，テント， 安心，ライト，ジムニー，納車，ワイン，レッスン， ヨガ，ツーリング，美品，ビンディング，コンサート， シリーズ，釣り，登山，クラブ，補修，強度，練習，ゲーム

また,「ブランド，予算，リクエスト，好み，こども，大人」など，選ぶ際のポイントや意識される要素が示されている.「実用品・普段使い」 のカテゴリでは,「タオル，靴下，シャンプー，財布，布団，毛布，シャツ，弁当」など，日常的に使うアイテムに関連する単語が多く含まれている．また,「はき心地，付け心地，締め付け，使い心地，指通り，着心地，音質，品質」など，使用感や快適さに関する単語や「毎月，継続，ストック，リピーター，まとめ買い，予備」など，定期的な購入や長期利用を前提とした単語が多いことがわかる.「趣味」のカテゴリでは,「ヨガ，ツーリング，ビンディング，コンサート，釣り，登山，ゲーム，プレイ，ワイン，クラブ」など，アウトドアやエンターテインメントなど，趣味に関連すると思われる単語が多く含まれている．また,「長持ち，補修，強度」など，趣味の道具を長く使うことや状態を維持することに関する単語も特徴的である．

楽天市場の「使い道」カラムのレビュー文中の単語の特徴として,「仕事関係へ」 のカテゴリでは,「上司，職場，会社，他社，得意」など，仕事の関係者向けの贈り物であることを示す単語が多く含まれている．また,「退職，定年，納会，忘年，イベント」など，職場のイベントや節目に関連する単語や「景品，評判，フルーツ，ビール，刻印」など，贈り

物として選ばれるアイテムに関する単語も含まれている.「友人へ」 のカテゴリでは,「友人，知り合い，同僚」など，友人関係を示す単語が登場している．また,「結婚，習い事，送り先」など，特定の交流の場でのプレゼントが想定される単語や「人気，予算，ブランド，パーティー」など，見栄えや選びやすさを意識した要素が含まれていることが特徴的である.「女性（彼女・妻）へ」 のカテゴリでは,「札入れ，本革，アロマ，輝き，紫色，発色」など，高級感やデザイン性のあるプレゼントが重視されていると思われる単語や「セットアップ，暖房，小型」など，実用性も兼ね備えた商品の特徴を表した単語が含まれている.「子供へ」 のカテゴリでは,「こども，小２，小学，学年，部活」など，子供向けのプレゼントであることを示す単語や「ランドセル，通学，スイミング，吹奏，弁当，保育」など，学校生活や習い事に関連する単語が特徴的である.「家族へ」 のカテゴリでは,「両親，父親，実家，高齢」など，家族向けや贈る対象を示す単語や,「大晦日，元旦，来年，今年，帰省」など，年末年始の贈り物や帰省時の手土産を示す単語が多く見られる．また,「リクエスト，珈琲，焼き肉，蕎麦，好物，すき焼き，数の子」など，家族が好む食品や飲み物が多く含まれていることも特徴的である.「男性（彼・夫）へ」 のカテゴリでは,「彼氏，職人，現場，運転」など，男性向けの贈り物であることを示す単語が登場している．また,「ベルト，ダウン，クロノグラフ，ウォーマー，アウター」など，ファッションや防寒具に関する単語が多く見られる.「自分用」 のカテゴリでは,「ｕｓｂ，インテリア，アイロン，ピアス，シャンプー，タオル，財布」など，自分用に購入する実用品や嗜好品に関連する単語が多い．また,「はき心地，付け心地，指通り，色合い，音質，着用，シルエット」など，使用感を重視する単語が多く含まれていることや,「旅行，通勤，普段，大量，収納」など，日常使いを想定した単語が特徴的である.「親戚へ」 のカテゴリでは,「叔父，叔母，祖父母，親族，義父，甥っ子，姪っ子」など，親戚向けの贈り物を示す単語が多く含まれている.「県外，田舎，郵便」など，年始の挨拶や遠方の親戚への贈り物として利用されることも示唆されている．上記の結果から，小川らと同様に，カテゴリごとのレビュー中の出現単語の特徴を把握することができたため，メタデータが完全に一致しなくても，一致条件を緩和するすることで代替できる可能性のあるデータカラムを推測することができるようになると考える．

表 5.7: 楽天市場の目的に特徴的なクラスタ群の一例

目的	クラスタのキーワード
仕事関係へ	共同，備品，名札，皆んな，箱入り，会社，他社，納会， 退職，定年，職場，ビール，景品，会場，刻印，忘年， 評判，フルーツ，男性，上司，イベント，年配，得意
友人へ	習い事，送り先，冷え性，レッスン，結婚，女性，ゴルフ， 人気，知り合い，予算，音色，同僚，日頃，見栄え，挨拶， パーティー，名入れ，赤ちゃん，友人，ブランド，バック
女性（彼女，妻）へ	札入れ，紫色，かみ，セットアップ，入園，レゼント 絵柄，本革，ママ，小型，アロマ，発色，暖房，輝き
子供へ	アレルギー，こども，保育，弁当，サンタ，部活，入学， チョコ，体型，通学，ランドセル，小２，吹奏，スイミング， 初詣，学年，冬休み，小学，シール，トランポリン，ゲーム
家族へ	両親，実家，大晦日，父親，リクエスト，珈琲，好物，蕎麦， 寝返り，人数，焼き肉，味噌，来年，今年，便利，高齢， 地元，数の子，帰省，すき焼き，評価，元旦，安心，毎年
男性（彼，夫）へ	ベルト，汚れ，ダウン，彼氏，不評，職人，現場，中敷き， クロノグラフ，イビキ，チャコール，奮発，ステーキ，現品， 運転，ウォーマー，アウター，太もも，ディスク，風呂上がり
自分用	usb，ジムニー，はき心地，肩幅，付け心地，指通り，毎月， 毛穴，財布，色合い，着用，純正，ワイン，化粧，季節，旅行， 音質，シルク，継続，インテリア，ストック，衣類，重宝，洗濯， シャンプー，タオル，色味，通勤，シルエット，大量，文句， 普段，受注，締め付け，予備，ピアス，納品，小分け，収納
親戚へ	ティー，定番，しゃぶしゃぶ，贅沢，集まり，返し，親族， 叔父，叔母，祖父母，羊羹，カステラ，義理，県外，甥っ子， 郵便，チョコレート，義父，内祝い，田舎，姪っ子

6　議論

本研究では，ユーザに新規データセットに適用可能な技術や得られる結果を推薦するために，既存の活用事例を「データ × 技術 × 結果」の形式で整理・分類する一端として，データの類似度を推定する手法を提案した．本章ではデータカラムの類似度推定を定性的に確認することで，提案手法が妥当であったか議論する．

6.1　メタデータによる類似カラムの推定

4，5 章で活用事例にて使用されたデータカラムから類似するデータカラムを推定した．推定したデータカラムが類似しているか，実データをもとに定性的に確認する．

6.1.1　実験例 1

4 章でメタデータによって類似すると判断されたデータカラムを使って既存の活用事例と同様な結果が得られたことから，メタデータによってデータの特性が類似しているものを推定できたと言える．小林らの使用した「評価ポイント」と「レビュー内容」と類似していると判断したデータカラムの例を表 6.1 に記す.「評価ポイント」は楽天市場に出品されている商品に対してユーザが 5 段階で評価した点数である.「ユーザ評価の評価 7（総合）」は楽天トラベルに掲載されている宿泊施設に対してユーザが総合的に 5 段階で評価した点数である.「総合評価」は楽天 GORA に掲載されているゴルフ場に対してユーザが総合的に 5 段階で評価した点数である．このような特性から,「評価ポイント」,「ユーザ評価の評価 7（総合）」,「総合評価」はデータ特性が類似していることがわかる．

「レビュータイトル」は楽天市場に出品されている商品に対してユーザが評価したデータであり,「かわいいデザイン」や「いまいち」など，ユーザの主観的な文章である.「レビュー内容」は楽天市場に出品されている商品に対してユーザが評価したデータであり,「良かった」や「辛い」といったユーザの主観的な文章である.「ユーザ投稿本文」は楽天トラベルに掲載されている宿泊施設に対してユーザが評価したデータであり,「利用したい」や「快適」など，ユーザの主観的な文章である.「コメント」は楽天 GORA に掲載されているゴルフ場に対してユーザが評価したデータであり,「ゆっくり」や「難しかった」などのユーザの主観的な文章である．これらのデータカラムは全て主観的な文章であるため，内容的特徴は類似しているが,「レビュータイトル」だけ文章が短いという特徴がある．

6.1.2　実験例 2

5 章でメタデータで類似すると判断されたデータカラムを使って既存の活用事例と同様な結果が得られたことから，メタデータによってデータの特性が類似しているものを推定できたと言える．小川らの使用した「性別」と「レビュー内容」と類似していると判断したデータカラムの例を表 6.4 に記す.「レビュー内容」は楽天市場に出品されている商品に対してユーザが評価したデータであり,「気分悪い」や「気に入っている」など，ユーザの主観的な文章である.「ユーザ投稿本文」は楽天トラベルに掲載されている宿泊施設に対してユーザが評価したデータであり,「良かった」や「利用したい」など，ユーザの主観的な文章である．このような特性から,「レビュー内容」,「ユーザ投稿本文」は主観的な文章であるという特性が類似している．

表 6.1: 楽天市場の例

評価ポイント	レビュータイトル	レビュー内容
5	かわいいデザイン。	三歳になる娘が、プール遊びやお風呂で大喜びで使ってます。 中身に水が入るのが見えるので、楽しのですかね。
3	いまいち	ワンサイズ大きめで買ったのですがきついです。デザインなどはとても良かったです。
5	これ大好きです	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。

表 6.2: 楽天トラベルの例

ユーザ評価の評価 7（総合）	ユーザ投稿本文
4	ご主人の応対と、掛け流しの温泉がとにかく良かった。泉質は、温泉街の共同浴場より良いくらい。 また利用したいと感じさせてくれる宿でした。
4	一昔前の保養所といった感じでした。これといった特徴はありませんが、 スタッフの皆さんの感じも良く、清掃も行き届いていて、快適に過ごせました。
5	久しぶりに宿泊しました。朝食付きとなり以前よりサービスが良くなりました。 これからも是非、利用したいと思います。

表 6.3: 楽天 GORA の例

総合評価	コメント
5	短いコースですが、樹齢の高い木々や自然のままのアンジュレーションを活かしたフェアウェイにより、 高い戦略性を要求されます。この価格設定で赤星四郎のコースが味わえるのはここだけでしょう。 レストランでも、生姜焼きが美味しく、付録の豚汁も浦霞の経営だけに（？）酒粕を使ったようで味わい深かったです。
3	天候にも恵まれたし、比較的空いていた事もありゆっくりラウンドする事ができました。
4	グリーンは凄く手入れされてて綺麗で、気持ち良かったんですが、私的には難しかったです。 また挑戦したいですね。お魚の煮付けと豚汁美味しかったです。 食事は、いろんなゴルフ場に行きましたが、ピカイチですね。

「使い道」は「イベント」や「プレゼント」など，楽天市場に出品されている商品をユーザがどのような使い道に使用するために購入したかを表すデータである.「目的（市場)」は「自分用」や「家族へ」など，楽天市場に出品されている商品を誰に向けて購入したかを表すデータである.「頻度」は,「はじめて」や「リピート」など，楽天市場に出品されている商品を何回購入したかを表すデータである.「目的（トラベル)」は「ビジネス」や「レジャー」など，楽天トラベルに掲載されている宿泊施設をユーザがどのような目的で利用したかを表すデータである.「同伴者」は「一人」や「家族」など，楽天トラベルに掲載されている宿泊施設をユーザが誰と利用したかを表すデータである．これらは，それぞれの要素ごとにデータを分類することができる特性を持つカテゴリデータである．小川らが使用した「性別」には「男性」と「女性」というカテゴリデータが存在したと推測されため，形式的な側面において「性別」と類似するデータカラムとして抽出できた．しかし，内容的な側面においては，類似しているとは言えない．

6.2 メタデータの妥当性

付与したメタデータがデータカラムの特性を適切に表現できているかについて検討する．4 章で紹介した小林らの研究では,「数値（順序)」と「指標データ（定性)」,「テキスト（文章)」と「指標データ（定性)」のメタデータを用いて，類似したデータカラムを算出した．

表 6.4: 楽天市場の例

使い道	目的	頻度	レビュー内容
実用品・普段使い	家族へ	はじめて	他店で同じ個数を購入した時は、新品が二箱、きれいな状態で届いたので、同じだろうと思って注文してしまいました。 よくレビューを読むべきでした。クッション材もなく、バラバラでダンボール箱に商品がそのまま詰められて届きます。 蓋のところがボコボコになっているのもありました。やはり見た目も気分も悪いので、こちらでは二度と注文しないでしょう。
イベント	自分用	はじめて	幼稚園でミシンが必要になるので購入しました。ミシンのケース付きでデザインがシンプル、 初心者が使いやすいのと白は限定というのにひかれ選びました。これから使うのが楽しみです。
実用品・普段使い	家族へ	リピート	リピートです。 合わせやすいデザインでとても気に入っています。
プレゼント	子供へ	はじめて	三歳になる娘が、プール遊びやお風呂で大喜びで使ってます。 中身に水が入るのが見えるので、楽しのですかね。

表 6.5: 楽天トラベルの例

目的	同伴者	ユーザ投稿本文
ビジネス	一人	出張で初めて利用しました。最初のみ東日本橋駅から少しだけ迷いましたが、駅からは近く、部屋も綺麗で快適に過ごせました。 これからも、利用したいと思います。部屋のアメニティーに、ボディータオルと綿棒があれば、さらに良かったと感じました。
レジャー	家族	新しく綺麗なホテルでまた、立地も良いので何度か利用させていただいています。全体的なコストパフォーマンスからみれば申し分ないです。 特に朝食サービスは今回もいただきましたが、味、品数ともに大満足です。しかし、以前から気になっている点は、 テレビの岩手局の放送が入らないことです。他の近隣ホテルでは視聴可能の所が多いので、検討して欲しいと思います。
レジャー	一人	以前より何度か利用させていただいています。やはり風呂好きの自分としては、大浴場があること。 また、朝食が美味しくてこのホテルを選択しています。また、利用したいと思います。
レジャー	恋人	部屋の眺望は全くだめ。最上階なのに目の前のビルしか見えません。前日お台場で宿泊していたので、その落差が激しすぎました。 ホテルまでのロビーやエントランスがものすごく都会的で、重厚感があるにもかかわらずお部屋は狭く窓も小さいです。 おまけに二人で宿泊したのに一人分のアメニティしか用意されてないし、フロントはすぐに対応してくれたのですが、印象が悪すぎです。

実験結果から，これらのメタデータの妥当性は確認されたものの，6.1 節でも述べたように，レビュータイトルは他の三つのデータカラムと比べて文字数が少ないという特性あるため，文字数に関するメタデータが必要である．

また，5 章で紹介した小川らの研究では,「テキスト（文章)」「記述（目的)」「指標（定量(回数))」のメタデータを用いて類似したデータカラムを算出した．実験結果から,「テキスト（文章)」のメタデータの妥当性は確認されたが，小川らの研究のように，レビューの傾向を属性ごとに把握することを目的とする場合，その目的に応じて分析対象となる文章を適切に絞り込むためのメタデータが必要であると考えられる．例えば，ユーザごとの特徴を分析する場合には,「ユーザ」（コンテンツを利用するユーザが作成した文章）や「会社」（コンテンツを提供する企業が作成した文章）といったメタデータを付与することで，より適切な分類が可能となると考えられる．

一方で，現状のメタデータを用いた場合，楽天市場における「使い道」と「目的」，楽天トラベルにおける「目的」と「同伴者」は「記述（目的)」のメタデータによって推定され，楽天市場の「頻度」は「指標（定量（回数))」のメタデータによって推定されている．また，6.1.2 章で述べたように，これらはカテゴリデータであり，これを推定可能なメタデータを追加することにより，効率的に類似するデータカラムを推定することができると考えられる．カテゴリデータのメタデータを自動付与することが可能になれば，異なるデータセット間でも代替可能なデータカラムの推定が実現できる可能性がある．

また，内容的側面について，楽天市場データセットは EC サイトの商品に対して付与されたデータセットであり，楽天トラベルデータセットは宿泊施設に対して付与されたデータセットであるため，それぞれのデータセットの対象が異なる．データの対象や観点を表すメタデータ（e.g.， 料理，サービス）を付与することで，メタデータを使用してより精度の高いデータの類似度を推定できる可能性がある．

また，現在は手動でメタデータ付与を行っているため，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する必要がある．

6.3 データ・技術・結果の三要素による体系的整理の効果

4章，5章より，メタデータによって推定された類似データカラムは，比較元のデータカラムと代替可能であることが明らかとなった．同様に，データカラムのみならず，技術や結果においても類似度を測定することが可能になれば,「データ × 技術 × 結果」の形式において，一要素が欠けた場合であっても，他の二要素の情報を用いて補完できると考えられる．今後は，ユーザの保有するデータ（e.g.， レビューデータ，評価データ）とユーザの目的（e.g.， 評価値を予測したい）をシステムに入力した際に，分析技術（e.g.， BERT，回帰分析）と具体的な結果（e.g.， 連続数値，離散スコア）を出力可能となれば，得られたデータの具体的な活用方策について明確な指針が得られていないユーザが既存の事例の活用方法を自身のデータに適用することができると考えられる．

本稿では人手によって付与したメタデータのみで類似度を推定したが，BERT などの手法によって類似度算出に重みづけができる可能性が考えられる．また，本研究では使用しなかったが,「良い」「普通」「悪い」や「とても満足」「どちらでもない」「とても不満」といったデータをスコアに変換することができれば，さらなる分析が可能となると考えられる．加えて，技術データベースや結果データベースにもメタデータを付与することで，より効果的な分析が実現可能となることが期待される．

さらに，データの組み合わせに関しても考慮する必要がある．同一データセット内の組み合わせでなければ意味を持たない場合がある一方で，異なるデータセット間での組み合わせによって有益な知見を得られる可能性もある．本手法によって，メタデータを選択することで類似したデータカラムを容易に特定することができた．

7　結論

本研究では，ユーザのデータ利活用を支援するために，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をもとにユーザの保有するデータに適用可能な技術を推薦する方法を示した．データセットの類似度を測る手法として，データカラムごとに形式的側面と内容的側面に着目したメタデータを付与し，メタデータの一致するデータカラム同士は類似度が高いものとした．

提案手法の有効性を検証するために，メタデータによって類似すると推測された楽天市場データセットを用いてレビューテキストからレビューの評価値を予測する活用事例を再現した結果，過去の事例より精度が良かった．これにより，本手法が一定の有効性を持つことを確認した．しかし，本研究では自分が閲覧可能なデータセットのみに限定されており，評価対象が楽天関連のデータセットに偏っているため，今後は異業種のデータへの適用可能性を検証する必要がある．さらに，多種多様な分析手法での検証が十分でない点も課題として挙げられる．例えば，ランダムフォレストやディープラーニングなど，他の分析技術での適用可能性について検討することは，技術選択の汎用性を確認する上で重要である．また，現在は手動で行っているメタデータ付与のプロセスについても課題があるため，メタデータ付与の自動化を進める具体的な方法を検討する必要がある．

今後は，ユーザの保有するデータとユーザの漠然とした目的をシステムに入力した際に，分析技術とその具体的な結果の一覧をシステムが出力するようにしていきたい．そうすることで，自身の保有するデータの活用方法について具体的な指針が得られていないユーザが，既存事例の活用方法を自身のデータに適用可能か把握できるようになり，業務の効率化に貢献するだろう．

謝辞

本研究をまとめるにあたり，関西大学総合情報学部の松下光範教授には数々のご指導ご鞭撻を賜りました．学会発表の際には，原稿や資料の添削をしていただき，より良い発表となるようご助言を賜りました．また，就職活動に際しては，面接内容や配属部署について親身に相談に乗っていただき，大変心強く感じました．私自身，説明が得意ではなく，拙い表現をすることが多々ありましたが，先生は私が伝えたい意図をしっかりと汲み取ってくださいました．そのおかげで，自分の考えを整理し，伝える力を養うことが少しはついたのではないかと思います．先生のご指導のもとで研究に取り組んだ 4 年間は，私にとって大きな成長の機会となりました．常に学生に対して真摯に向き合い，ご指導いただいたことに深謝いたします．

関西大学大学院総合情報学研究科知識情報学専攻の松下研究室所属の 13 期生の皆様に感謝申し上げます．大学院時代には同期に女子がいなかったため，最初は心細い気持ちもありましたが，みなさんが温かく接してくれたおかげで，楽しく過ごすことができました．特に，畑玲音氏には深く感謝しております．研究に行き詰まったときには，何度も相談に乗っていただき，的確なアドバイスをいただきました．また，研究室での何気ない雑談にも付き合ってくれたことが，日々の研究の合間の良い息抜きになりました．ほぼ毎日のように学校に来ていたため，研究室に行けば畑氏がいるという安心感があり，とても心強かったです．また，新川晴紀氏とは，共に学会発表に挑戦し，発表練習を重ねたことは，貴重な経験となりました．また，研究室での雑談にも付き合っていただき，研究の合間にリラックスすることができました．畑氏，新川氏とは特に長い時間を共に過ごしました．日々の研究や発表準備だけでなく、何気ない会話や休憩の時間も含めて、3 人で過ごした時間はとても楽しく、かけがえのない思い出となりました．田中瑠彗氏には，就職活動の際に，親身になって相談に乗っていただきました．また，志望していた企業に落ちた際には，温かい言葉をかけていただき，大きな励みとなりました．

研究生活を送る上でお世話になりました 11 期生，12 期生，14 期生，15 期生の皆様に深く感謝申し上げます．特に，中学時代からの友人である森岡真由氏には感謝の意を表したいと思います．もし森岡氏がいなかったら，そもそもこの大学を受験するという選択肢を考えることすらなかったかもしれません．同じ研究室に所属し，さらに中高時代から共に過ごしてきたからこそ，気兼ねなく話せることも多く，研究の合間に交わした会話はとても楽しいものでした．忙しい日々の中で，森岡氏がよく晩御飯を作ってくれ，一緒に食事を楽しんだことも忘れられません．学部生の頃から院生となった今に至るまで，常にそばにいてくれたことに心から感謝します．この先も変わらず，どうぞよろしくお願いいたします．

最後に，6 年間という長い期間の学生生活を支えてくれた両親に心から感謝の意を記して謝辞と致します．

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」（https://rit.rakuten.com/data_release/）を利用した．記して謝意を表す．

A　付録

3.3 節で得られたデータセットのカラム特徴のうち，本文中に掲載しなかったものを以下に示す.

表 A.1: 「レシピ情報」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
レシピ 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
ユーザ 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
大カテゴリ	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
中カテゴリ	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
小カテゴリ	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
レシピタイトル	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
レシピのきっかけ	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
レシピの紹介	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
料理画像ファイル名	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
料理名	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

表 A.2: 「材料情報」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
レシピ 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
素材名	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
分量	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0

表 A.3: 「作り方情報」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
レシピ 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
手順位置	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
手順作り方	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

表 A.4: 「「つくったよ」レポート情報」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
レシピ 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
ユーザ 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
おすすめコメント	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
オーナーコメント	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
作成日時	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

表 A.5: 「Pickup レシピ」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
ピックアップ日	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
レシピ 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

表 A.6: 「デイリシャスニュース」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
ニュース記事 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
公開日時	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
ライター 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
ジャンル名	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
記事タイトル	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
記事テキスト（メイン記事）	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

表 A.7: 「ユーザ評価」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
投稿日時	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
施設 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
プラン 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
プランタイトル	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
部屋種類	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
部屋名前	1	0	1	0	0	0	0	0	1	0	0	0	0	1	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
同伴者	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
評価 1（立地）	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
評価 2（部屋）	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
評価 3（食事）	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
評価 4（風呂）	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
評価 5（サービス）	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
評価 6（設備）	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
評価 7（総合）	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
ユーザ投稿本文	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
施設回答本文	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

表 A.8: 「施設名マスタ」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
施設 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
施設名	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

表 A.9: 「クチコミ情報」に付与したメタデータ

	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
クチコミ 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
コース 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
クチコミ投稿者名	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
都道府県	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
年齢	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
平均スコア	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
オススメ目的	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
オススメタイプ	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
利用回数	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
総合評価	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
コストパフォーマンス	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
スタッフ接客	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
コース/戦略性	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
食事が美味しい	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
設備が充実	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
フェアウェイが広い	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
距離が長い	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
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
コメント	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
クチコミ登録日	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
プレー日	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

参考文献

[1] Arthur, D. and Vassilvitskii, S.: k-means++: The Advantages of Careful Seeding, *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035 (2007).

[2] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2019).

[3] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A. H.: *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute (2011).

[4] Masaaki, S., Naoki, S., Soma, K. and Fumiaki, S.: An Analysis on the Customer Logistic Satisfaction based on Word Clustering, *Innovation and Supply Chain Management*, Vol. 17, No. 1, pp. 11–16 (2023).

[5] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *Proceedings of the 1st International Conference on Learning Representation*, arXiv:1301.3781v3 [cs.CL] (2013).

[6] Murray-Rust, P.: Open Data in Science, *Nature Precedings* (2008).

[7] Ohsawa, Y., Kido, H., Hayashi, T. and Liu, C.: Data Jackets for Synthesizing Values in the Market of Data, *Procedia Computer Science*, Vol. 22, pp. 709–716 (2013).

[8] Ruijer, E.: Designing and implementing data collaboratives: A governance perspective, *Government Information Quarterly*, Vol. 38, No. 4 (2021).

[9] Taichi, N., Atsushi, H., Yoshitaka, U. and andMori Shinsuke, K. H.: State-aware Video Procedural Captioning, *Multimedia Tools and Applications* (2023).

[10] Yuanyuan, W.: Hotel Recommendation based on Review Analysis by Considering Tourism Needs, *Proceedings of the 16th International Conference on Ubiquitous Information Management and Communication (IMCOM 2022)*, pp. 1–8 (2022).

[11] 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).

[12] 大須賀智子, 大山敬三: 情報学研究データリポジトリ IDR における研究用データセット共同利用の取り組み, 情報処理学会論文誌デジタルプラクティス, Vol. 2, No. 2, pp. 47–56 (2021).

[13] 大須賀智子, 大山敬三: 国立情報学研究所「情報学研究データリポジトリ (IDR)」の紹介, 自然言語処理, Vol. 29, No. 3, pp. 986–990 (2022).

[14] 大山敬三, 大須賀智子: 国立情報学研究所における研究用データセットの共同利用, 情報管理, Vol. 59, No. 2, pp. 105–112 (2016).

[15] 小川和晃, 田村哲嗣, 速水悟: 商品レビューにおける時系列情報に着目したクラスタ分析と可視化, 人工知能学会全国大会論文集, No. 1F5-5 (2016).

[16] 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).

[17] 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).

[18] 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).

[19] 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).

[20] 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).

[21] 総務省: 第2章 データの流通・活用の現状と課題, 令和5年版 情報通信白書 第1部 (2023).

[22] 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).

[23] 早矢仕晃章, 大澤幸生: データの価値と利用方法発見のための創造的コミュニケーションとメタデータ記述方法の提案, 人工知能学会全国大会論文集, No. 2F1-2 (2014).

[24] 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).

[25] 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).

[26] 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).

[27] 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).

[28] 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).

[29] 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).

[30] 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).