学術論文の特徴を用いたデータセット特性の表現に関する研究

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

20M7102
玄道 俊

論文要旨

1　はじめに

共同利用可能なコンテンツに関するデータセット（コンテンツデータセット）は，様々な研究への利用を念頭に公開されており，これらのデータセットを利用した研究が数多く報告されている．現状では，コンテンツデータセットは，扱うコンテンツの違い（例えば，レシピやホテルレビューなど）に従ったデータ項目で整理されている．そのため，異なるデータセット間では，共通あるいは類似した性質をもつデータ項目が存在していたとしても，それらの関係性を読み解くことは容易ではない．この問題を解決するためには，複数のコンテンツデータセット内のデータ項目を横断的に整理可能にする必要がある．本研究では，複数のコンテンツデータセットに含まれるデータ項目群を横断的に理解するための情報整理フレームワークを提案する．提案手法では，コンテンツとコンテンツを提供するクリエイター，コンテンツを利用するユーザの 3 者の関係によって抽象化しその妥当性を検証する．

2　提案手法

異なるコンテンツデータセット間のデータ項目において，意味的な共通性が認められるデータ項目が存在する．しかし，データ項目の意味的な共通性は表層的ではなく，データ項目の意味的な共通性に着目して抽象化することで複数のコンテンツデータセット内のデータ項目を横断的に整理する必要がある．意味的な共通性を持つデータ項目を利用した研究では，目的・課題の類似性や用いている手法に類似性が見られる可能性がある．コンテンツを説明する項目を対象とした研究は，そのコンテンツの特性の解明に主眼があるのに対し，人がコンテンツに関わることで生み出された項目を対象とした研究は，コンテンツ自体にとどまらず，コンテンツの利用者や制作者の意図や特性の解明などを射程に入れている．

本稿では，「人（クリエイターとユーザ）とコンテンツの関係」に着目し，RDF 形式で記述し有向グラフ化することで抽象化を行う．RDF 形式では「＜主語＞は＜目的語＞を＜述語＞する」の関係でデータ項目間の関係性を表現する．＜主語＞を起点ノードに設定し，＜目的語を＞終点ノードに設定する．これをコンテンツデータセットの全てのデータ項目に適用することで，コンテンツデータセットにおけるデータ項目の関係が概念的に表現される．さらに，複数のコンテンツデータセットの共通性を見ることで，複数のコンテンツデータセットを横断的に把握することが可能になる．

3　対象とするコンテンツデータセットの抽象化

対象とするコンテンツデータセットは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，クックパッドデータセットの 4 種類とした．有向グラフ化にあたり，ノードにはコンテンツ・クリエイター・ユーザの 3 者を設定し，エッジにはこれら 3 者の関係性を表す述語を設定した．エッジとなる述語については，目的をメタ的に説明する「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 種類とした．以上に従って，コンテンツデータセットを有向グラフ化し抽象化した結果の一例を図 1 に示す．データ項目は全 7 種類の RDF 形式に分類された（以下，共通項目と記す）．7 種のデータ項目の分類に該当する，各コンテンツデータセットの各データ項目の例を表 1 に示す．

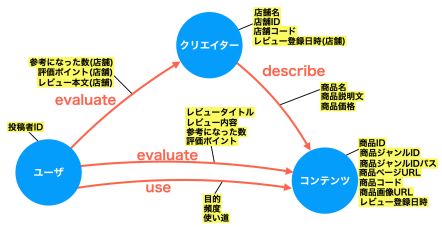

図 1: 有向グラフ化の一例：楽天市場

表 1: RDF 形式の共通データ項目（一部抜粋）

	主語	目的語	述語	楽天市場	楽天トラベル
Ⅰ	ユーザ	コンテンツ	evaluate(離散)	評価ポイント 参考になった数	評価 1,2,3,4,5,6,7
Ⅱ	ユーザ	コンテンツ	evaluate(text)	レビュータイトル レビュー内容	ユーザ投稿本文
Ⅲ	ユーザ	コンテンツ	use	目的 頻度 使い道	目的 同伴者

4　抽象化による共通項目の妥当性

提案フレームワークの妥当性の評価は，「ある共通項目を用いた論文は，その論文と同じ共通項目かつ別のコンテンツデータセットを用いて研究している論文と類似している」という仮定のもと定量的な検証を行い，実データの中身を確認することで定性的な検証を行った．検証に利用した論文は，国立情報学研究所の情報学研究データリポジトリから各コンテンツデータセットにつき 25 本，合計 100 本である．論文間の類似度算出は，その論文のコンテンツに依存した単語と，全ての論文で文書頻度数の高い単語の除去を行った後，対象論文の出現単語から作成した度数分布に対して Bhattacharyya 距離を適用することで行った．ある共通項目を用いた論文は，同一の共通項目かつ別のコンテンツデータセットを用いた論文の Bhattacharyya 距離の平均順位は平均 1.88 位となり，定量的な検証では提案フレームワークの妥当性が示された．

次に，定性的な検証をするために，各共通項目の実データの確認を行った．共通項目Ⅲのクックパッドの「つくれぽ内容」には「ワカモレ美味しく出来ました！」，楽天レシピの「おすすめコメント」には「簡単にできました！」と記述されており，両者はともに「“ユーザ” が “コンテンツ” を “use” する」という構造が確認できた．しかし，共通項目Ⅲの楽天市場の「使い道」には【趣味】や【イベント】といった内容が記述され，楽天トラベルの「同伴者」の項目には【一人】や【家族】といった内容が記述されていた．これらの項目は単語のみの記述であり，RDF 形式の共通項目Ⅲには該当しない．定性的な調査の結果，共通項目Ⅱ, Ⅳ, Ⅶに分類されたデータ項目の実データは，それぞれの「＜主語＞が＜目的語＞を＜述語＞する」の関係であった．そのため，これらの共通項目においてコンテンツデータセット間の関係性を横断的に把握することが可能であった．ただし，一部のデータ項目は，どの共通項目にも該当しなかったことから，その関係性を横断的に把握することはできなかった．

5　おわりに

本研究では，複数のコンテンツデータセットに含まれるデータ項目群を横断的に理解するための情報整理フレームワークを提案し，その妥当性を定量的かつ定性的に検証を行った．今後，複数の共通項目に該当するデータ項目の検証を行い，提案したフレームワークの精度向上を目指す．

参考文献

[1] Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions, *Bulletin of the Calcutta Mathematical Society*, Vol.35, pp. 99-109 (1943).

目 次

1　序論

本章では，本研究の実施に至った背景を説明し，対象とする課題を明確にする．

1.1　データ収集と利活用の現状

IT 社会を生きる我々はデータの蓄積と利用の恩恵を受けている．例えば，ユーザが保有する携帯端末に搭載されている位置情報から旅行者の興味を引く対象を推定したり [15]，不動産情報サイトに掲載された物件データから場所や間取りを考慮し物件の価格相場を算出するアルゴリズムの開発 [29] を行うことで，実際のサービス開発に繋がっている．そのため，サービスの開発や提供を実現させるには，企業はデータの利活用が求められ[1]，データの利活用は労働や資本と同様に重要な要素の 1 つとなっている [8][18].

データの利活用を行うためには，データの収集・蓄積から可視化といった分析や予測する必要がある．企業はデータの利活用を行う過程で，例えば既存のサービスの発展や新規サービスの参入といったビジネスモデルの転換が行われ，データの利活用は新たな価値を生み出す可能性がある．しかし，総務省の調査によると，日本における企業のデータ収集・蓄積に取り組む企業は 51.5%となり，データ分析の結果を活用し業務効率が向上した企業は 22.5%に留まった．加えて，新たなビジネスモデルによる付加価値の拡大をした企業は 13.4%に留まった（図 1.1 参照）.

企業がデータを活用できない理由として，多くの企業は社内に十分な研究開発能力を備えていないことがあげられる [20]．そのため，経済産業省は産学連携推進事業費補助金[2]を出すといった取り組みを通して産学連携を推進し，大学の教育機関や研究機関と共同でサービスの開発やデータの分析を行うことで，産業の活性化を促している．

1.2　研究機関のデータ利活用について

データ利活用可能な技術をもつ研究者がデータを用いて研究を行うためには，目的に応じたデータを効率よく収集し，データセットを構築する必要がある．研究者にとってデータを収集し構築することが困難な理由として，大山らは以下の 2 点を指摘している [20].

- コストがかかりすぎることにより，十分な規模が確保できない.
- 偏りなく収集することが難しい.

この 2 つの問題を解決する方法の 1 つとして，計算機によってデータを収集することが考えられる．Web 上に存在する民間企業のサービスに関わるデータをスクレイピング等の技術を用いて収集し，そのデータからデータセットを構築する．しかし，これらの技術を使用しデータセットを構築する際にも，以下の問題点があると指摘されている.

- データセットの構築方法や利用方法の適法性が 明確でない場合がある.

[1]総務省 平成 28 年版 情報通信白書: https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/html/nc123200.html(2022/02/06 確認)

[2]https://www.meti.go.jp/information/publicoffer/kobo/2022/k220120002.html（2022/2/13 確認）

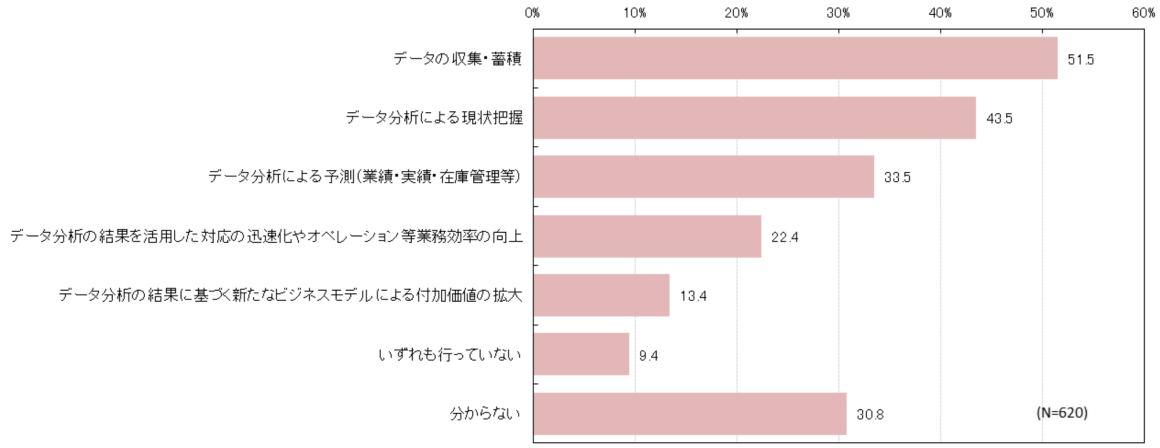

図 1.1: 日本企業におけるデータの利活用状況 (文献 [24] より図引用)

- データセット中の個々のデータに含まれる未処理の著作権や個人情報などが顕在化する．

これらの理由からデータセットの構築を躊躇する研究者も存在する [20].

1.3 共同利用可能なデータセットの活用

1.1 節で述べたデータセットを利活用できていない企業と 1.2 節で述べデータセット構築するのが困難な研究者を結びつける 1 つの方法として，共通のライセンスで利用できるデータセットの需要が高まっている．そのため，共同利用可能なデータセットを提供するレポジトリは年々増加しており，学術論文へのオープンアクセスを担ってきた機関レポジトリにデータを公開する企業なども増えてきている [16][11][13].

様々な目的に応じた研究利用を念頭に，様々なコンテンツに関するデータセット（以下，コンテンツデータセットと記す）は共同利用可能なデータセットとして公開されている．それらのコンテンツデータセットを利用した研究も増加しており [16][14]，国立情報学研究所が提供するコンテンツデータセットを用いた研究は，2022 年 1 月の時点で 1,143 件あると報告されている[3]．研究者は，共同利用可能なコンテンツデータセットを用いて研究を行う利点は，同じデータセットを利用した他の研究成果との比較・検証が可能になる点である [20]．また，研究者が新しく提案する手法を異なる複数のコンテンツデータセットに適用することができれば，その提案手法の汎用性を検証することも可能になる．

1.4 本研究が解くべき課題

コンテンツデータセットは，コンテンツの特性に基づいたデータ項目で構成されている．例えば，レシピデータのコンテンツデータセットであれば「材料」や「手順」「レビュー文」，ホテルデータのコンテンツデータセットであれば「宿泊の目的」や「同伴者」「ユーザー投稿本文」といった，それぞれのコンテンツを特徴化するデータ項目が存在する．レシピデー

[3]国立情報学研究所 HP: `https://www.nii.ac.jp` (2022/02/06 確認)

タのデータ項目である「レビュー文」とホテルデータのデータ項目である「ユーザ投稿本文」は,「ホテル」と「レシピ」という扱うコンテンツが異っているが,「コンテンツ」を「人」が評価するという点でこれらのデータ項目は同一な性質を持つと言える.

上述のように，こういったデータ項目が他のデータセットのデータ項目と同一の性質を持つことが明らかになれば，研究者は提案手法の汎用性を検証することが可能になる．しかし，実際にはデータセットが異なると，同一の性質を持つデータ項目であっても，それぞれが扱うコンテンツが異なることでデータ項目の名称も異なり，それらの関係性を読みとくことは困難である．それらの関係性を読み解くためには，異なるデータセット間のデータ項目を横断的に把握することが必要不可欠となる．

本稿では，複数のコンテンツデータセットに含まれるデータ項目群を横断的に理解するための情報整理のフレームワークを提案する．国立情報学研究所の情報学研究データリポジトリが提供するコンテンツデータセットを対象とし，コンテンツデータセットに含まれるデータ項目の性質を整理する．提案フレームワークでは,「データ自体の特性を表現するもの」と「そのデータがどのような目的で作られたのか」という2つの観点からデータ項目を整理して，データセットを概念レベルで共通化する．

2　関連研究

データセットの活用を促進するために，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，研究論文中でのデータセットの利用傾向の把握やデータセットの情報表現について概観し，本研究を位置づける．

2.1　データセットの利用傾向の把握に関する研究

研究論文の枠組みの中で，データセットの利用傾向の把握に関する研究としては，研究で用いられたデータセット名の論文から抽出が行われている．

Ayush らは，多種多様である研究課題に対して最も有用なデータセットを選択することが困難である問題を解決するために，研究論文から NGD (Normalized Google Distance) を用いてデータセット名を抽出している [12]．この手法では，学術的な検索エンジンが研究論文に関する情報を整理された形で提供していることに着目して，訓練データに依存することなく，また文書全体をスキャンすることなく自動的にマイニングする事が可能である．Ayush らの手法は，精度，再現率，F 値などの情報検索指標において良好な結果を示し，研究テーマごとに分類された研究論文のライブラリが整理されている条件のもと，研究で用いられたデータセット名を高精度に抽出できることを示した．

Behnam らは，社会科学の分野においてデータセットに言及しているにも関わらず，データセットへの明示的なリンクが提供されていないことを問題とし，半自動的にデータセットを見つける方法を提案している [5]．論文内の文を単語分割し，tf-idf とコサイン類似度を用い，データセットタイトルと類似度を図り，適切なデータセットをピックアップしてくる．このとき，”Study”, ”Survey”などのデータセットを参照する用語を手動で作成し，これらが論文内で出現した際は高い重みを与えることとした．図 2.1 の Paper1 内に出現した “study allbus 2014” は既存の tf-idf 法では “Allbus 2014” の類似度が高くなるが，提案手法により “Study Allbus2000” との類似度が高くなる．Behnam らの手法を「論文のデータセット参照を特定」と「論文で検出されたデータセット参照の項目と照合」の 2 つの観点から評価した結果，両者ともに良好な結果を得ることができた．

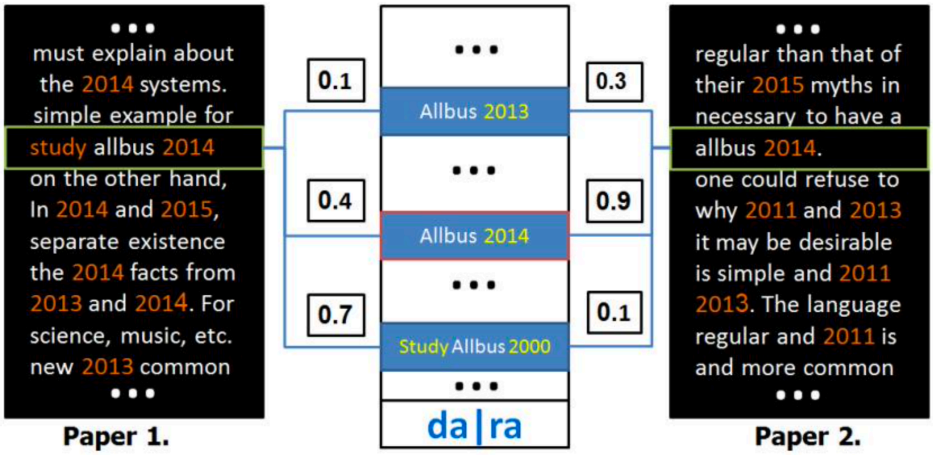

図 2.1: tf-idf の重み付け（文献 [5] より図引用）

また，Ikeda らは，学術論文から手法等を抽出し，二次利用を促進することを目的とし，その端緒として，データセット名の自動抽出を試みている [6]．word2vec で作成したモデルを用いて “dataset,” “datasets,” “database,” “databases” との類似度を測り, いずれかとの類似度が任意の閾値以上であった場合データセット名とした．データセット名の抽出精度の評価において，従来では分野ごとのデータセット名の辞書が必要であったのに対し，情報検索システムで用いられる尺度である precision@N と推薦システムで用いられる nDCG を用いることでデータセット名の辞書を必要としない定量的な評価を実現した．

本節の研究では論文からデータセット名の自動抽出を目指している．しかし，データセットの使われ方の特定，及び共通化までは踏み込んではいない．

2.2 メタ情報の付与を行ったデータセットを用いた研究

データセットにメタ情報を付与し，その情報を用いた研究が行われている．

Ohsawa らは，データ自体を秘匿したまま，データの概要情報 (e.g., データに関する説明文，変数名，保存形式）を記述するための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 [9][27]．データジャケットは機械の可読性を高めるためではなく，人間がデータについて理解することを目的としている．このフレームワークにより，実データの変数や値といった中身を公開することができずとも，データの所在と有している情報を理解することが可能となり，データの利用方法を検討可能にする（図 2.2 参照）．データ市場を模したワークショップである Innovators Markket Place on Data Jackets (以下，IMDJ と記す) がある [10][18]．IMDJ はデータジャケットを媒介し，データ利用者や提供者，分析者が要求を出し合うことで問題を顕在化して解決方法を提案していく場である．データ保有者はデータジャケットを IMDJ に提供することで，活用方法の可能性を知ることが可能である．

上原らは，データ保有者が IMDJ にデータ提供後にしかどのような課題解決に使われるかわからないことに着目し，データジャケットの類似度を図り類似したデータジャケットを提示するモデルの作成を行っている [17]．このモデルの類似度の指標はデータジャケットのタイトル，概要，変数名の 3 項目を用いている．これらの類似度と重み付けの定数を掛け合わせたものを，データジャケットの類似度とした．類似度は Bag of Words と tf-idf, 重み付けは重回帰分析を用いて算出している．評価の結果，上原らのモデルでは，概要の類似度評価に最も重み付けがされていることが明らかになった．

上原らの類似したデータジャケットを提示する手法ではデータセットと同様のコンテンツのデータセットを探索可能であるが，コンテンツが異なるものの提示は困難である．

2.3 データセットの情報表現についての研究

データセットの情報表現についての研究としては，データ項目間の関係性を述語で表現する研究が行われている．

久永らは，行政・団体がオープンデータを「データの 2 次利用が可能である」といった活用まで至ってない問題に対し，地方公共団体から収集した 626 個のオープンデータを RDF 形式へ変換している [28]．久永らの研究では，Resource Description Framework(以下，RDF

実データ

年	月	日	顧客ID	購入品目	支払金額(円)
2018	3	1	AAAAA	ファジィ学会誌, ビール	4567
2018	3	1	BBBBB	ペン, りんご, パイナップル	1080
⋮	⋮	⋮	⋮	⋮	⋮
2018	3	31	YYYYY	スナック, するめ, ビール	2536
2018	3	31	ZZZZZ	ラーメン, ナタデココゼリー	867

変数

値

データジャケット (*DJ*)

タイトル	××スーパーマーケットのPOSデータ
データ概要	××スーパーマーケットで収集されている顧客の購買行動履歴のデータ.
変数ラベル	年
	月
	日
	顧客ID
	購入品目
	支払金額
データのフォーマット	CSV
データの種類	テキスト
	数値
共有条件	共有不可
分析方法	時系列分析
分析結果	その日の売上の計算, トレンド商品の特定
分析方法以外に期待する分析・結果	顧客の購買行動とリピート率から, ロイヤルカスタマーを特定する.
データの所有者とその所在	××スーパーマーケット
データの収集方法・コスト	購入時に提示されたポイントカードの記録から取得
コメント(データに関する補足事項)	他データとの有用な組み合わせが発見できれば, データの提供, コラボレーションの可能性あり.

図 2.2: 実データのデータジャケット化と記述例（文献 [27] より図引用）

と記す) 形式[1]への変換を行うために，Word2Vec を用いて述語ベクトルの生成とクラスタリングを行った．これによりオープンデータが持つ項目を住所系，番号系，名称系，数値系，URL 系でクラスタリングすることを可能にした 2.3．久永らの研究では，行政・団体のオープンデータという同一種の大量のデータセットを対象としている．一方で，同一種のデータセットがクラスタリング可能なほど存在しているとは限らない．例えば，国立情報学研究所が提供しているコンテンツデータセットは 14 社 23 種類[2]となっており，異なる種類のデータセットが少数ずつ存在する．そのため，久永らの手法をそのまま適用して，データ項目の情報表現を行うことは難しい．

[1]W3C 公式 RDF サイト (https://www.w3.org/RDF/)

[2]国立情報学研究所 情報学研究データリポジトリ: 民間企業提供データ一覧: `https://www.nii.ac.jp/dsc/idr/datalist.html`(2022/02/09 確認)

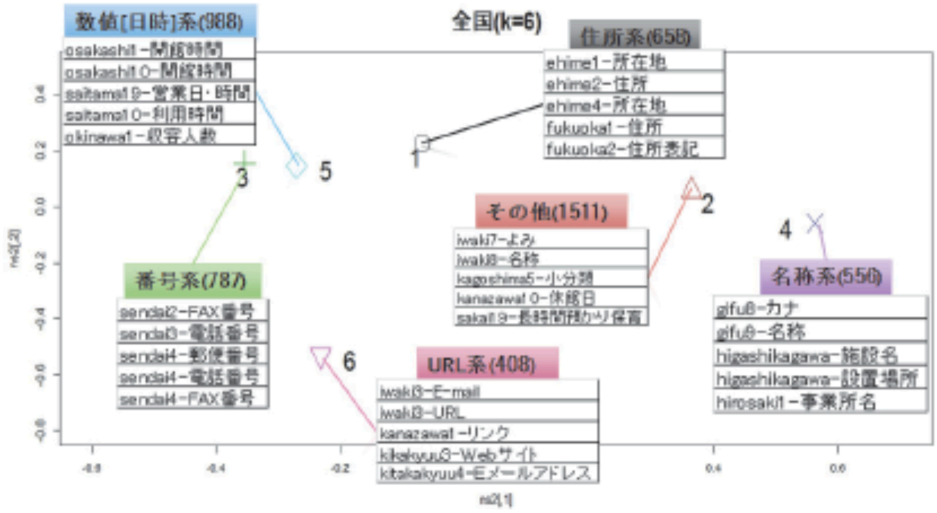

図 2.3: 行政・団体のオープンデータを用いたデータ項目のクラスタリング結果（文献 [28] より図引用）

3　提案手法

本章ではコンテンツデータセットの共通性について述べたうえで，抽象化の方法について述べる．

3.1　コンテンツデータセットの共通性

コンテンツデータセットは，コンテンツに紐付いた様々なデータの集合である．コンテンツデータセットには，自身で作成した料理のレシピの概要や材料，手順の情報や，宿泊したホテルのレビュー文，宿泊目的の情報など様々な種類が存在する．

これらの異なるコンテンツデータセットには，同じ意図や目的で作られているデータ項目が含まれる場合がある．実際にサービス展開されているものを例として挙げると，楽天市場[1]と楽天トラベル[2]の 2 つのコンテンツデータセット間において，コンテンツに対し「評価」する箇所で共通性が見られる．楽天市場の「この商品を購入された方のレビュー」欄には，実際に楽天市場を使用し商品を購入したユーザの評価に関する評価が文章として記述されている[3]（図 3.1 を参照）．また，楽天トラベルの「お客さまの声」欄には，実際に楽天トラベルを使用し宿泊した施設や利用しに旅行プランに対するユーザの評価が文章として記述されている[4]（図 3.2 参照）．これらのレビュー文は商品を扱ったユーザが投稿欄から投稿している（図 3.3 参照）．

コンテンツに関与する「人」の立場が異なれば意図や目的が異なるため，付与される情報も異なる．コンテンツに情報を投稿する「人」のケースで例をあげる．楽天市場の商品に対するレビュー文は「レビュー内容」，楽天トラベルの施設・プランに対するレビュー文は「ユーザ投稿本文」として，各コンテンツデータセットに格納されている．これらの「レビュー内容」と「ユーザ投稿本文」はどちらもコンテンツに対する人の評価であるという共通性が認められる．

クックパッドと楽天市場の 2 つのコンテンツデータセット間において，コンテンツに対し「説明」する箇所で共通性が見られる．図 3.4 はクックパッドのレシピ画面[5]，図 3.5 は楽天市場の商品ページである．図 3.4 の赤枠部分は料理レシピに対する説明を文章として記述している．同様に，図 3.5 の赤枠部分は商品に対する説明を文章として記述している．これら料理や商品といったコンテンツに対する説明文を人が投稿欄から投稿している (図 3.6 参照)．クックパッドの料理に対する説明文は「レシピの概要」，楽天市場の商品に対する説明文は「商品説明文」として，各コンテンツデータセットに格納されている．これらの「レシピの概要」と「商品説明文」はどちらもコンテンツに対する人の評価であるという共通性が認められる．

---

[1]楽天市場 HP: https://item.rakuten.co.jp(2022/02/03 確認)

[2]楽天市場 HP: https://travel.rakuten.co.jp(2022/02/03 確認)

[3]楽天市場にくらす スライドペールゴミ箱: https://item.rakuten.co.jp/nikurasu/641123481s/?s-id=ph_pc_itemname (2022/02/03 確認)

[4]楽天トラベル ANA クラウンプラザホテルグランコート名古屋: https://review.travel.rakuten.co.jp/hotel/voice/1659/?f_time=&f_keyword=&f_age=0&f_sex=0&f_mem1=0&f_mem2=0&f_mem3=0&f_mem4=0&f_mem5=0&f_teikei=&_version=2&f_static=1&f_point=0&f_sort=0&f_next=20(2022/02/03 確認)

[5]Cook Pad 簡単!定番 豚バラ白菜のミルフィーユ鍋: https://cookpad.com/recipe/2924774 (2022/02/09 確認)

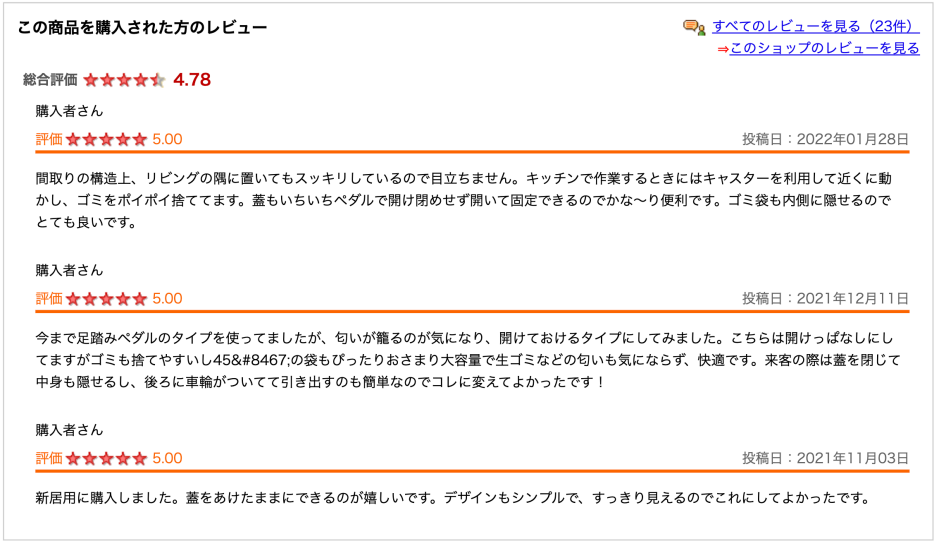

図 3.1: 楽天市場における商品の評価画面

上記の例のように，異なる種類のコンテンツデータセットであってもそこに含まれるデータ項目には意味的な共通性を持つ項目がある．こうした観点から，本研究ではコンテンツデータセットのデータ項目を，コンテンツ自身を説明する項目（e.g., 商品 ID，発売日時，画像 URL）と，人がコンテンツに関わることで生み出された項目（e.g., レビュー文，レシピのコツ）という 2 種類の意味的項目に大別する．

3.2　データセットの共通性を見つけるメリット

意味的に共通したデータ項目を利用した研究では，目的・課題の類似性や用いている手法に類似性が見られる可能性がある．コンテンツ自身を説明する項目を対象とした研究は，そのコンテンツ自体の特性の解明に主眼があるのに対し，人がコンテンツに関わることで生み出された項目を対象とした研究は，コンテンツ自体にとどまらず，そのコンテンツの利用者や制作者の利用意図及び，特性の解明などを射程に入れている．したがって，「どのような研究で利用されるデータ項目であるのか」によってデータ項目の性質を把握できる可能性がある．この性質を把握するためには，複数のコンテンツデータセット内のデータ項目を横断的に整理可能にする必要がある．

3.3　抽象化の基本的な考え方

複数のコンテンツデータセット内のデータ項目を横断的に整理可能にするためには，データ項目の意味的な共通性に着目して抽象化することが求められる．本稿では，意味的な共通性として「コンテンツと人の関係」を足がかりとしたデータ項目の抽象化を試みる．まず，コンテンツと関係する「人」についてより詳細化する．本稿では，コンテンツと関係する「人」を，コンテンツに情報を提供する人と，その投稿されたコンテンツを利用する人といった 2 つの立場に分けることで詳細化を行う．

投稿型料理レシピサイトの場合は，作った料理のレシピをサイトに提供する人と，その

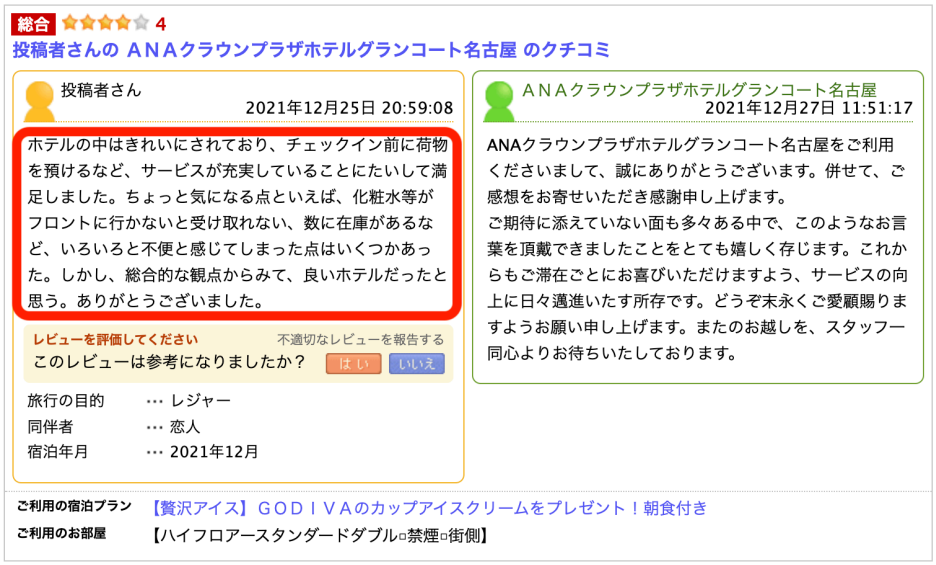

図 3.2: 楽天トラベルにおけるホテルの評価画面

レシピを利用して料理を作る人が存在する．レシピ提供者は，レシピサイトに作った料理を説明するために「レシピ概要」や「レシピのコツ・ポイント」などの項目を記載する．レシピ利用者は，そのレシピを評価するために「レビュー」を記載したり，「評価ポイント」を付与したりする．オンラインショッピングサイトの場合は，売却する商品をサイトに掲載する人と，その商品を購入する人が存在する．商品出品者は，ショッピングサイトに商品を説明するための「商品名」や「商品説明」,「商品価格」などの項目を記載する．商品購入者は，その商品や店舗を評価するために「レビュー」を記載する．

こうした観点に基づき，コンテンツデータセットに含まれる項目を，

(1) コンテンツ

(2) コンテンツを提供するクリエイター

(3) コンテンツを利用するユーザ

の 3 者の関係によって整理可能であると考え，これらの 3 者関係によってコンテンツデータセットのデータ項目を表現する．

3.4 RDF に基づくデータ項目の抽象化

本稿では，データに関する情報の記述方式の 1 つである RDF 形式を用いる．RDF 形式では，主語・述語・目的語の 3 つの要素でデータ構造を定義し [28],

> 「＜主語＞（ノード１）は ＜目的語＞（ノード２） を ＜述語＞（エッジ）する」

の関係でデータ項目間の関係が概念的に表現される．本稿では，コンテンツデータセットのデータ項目の関係性を，RDF 形式の有向グラフを用いて表現することで複数の異なるコンテンツデータセットのデータ項目に共通した抽象化を行う．有向グラフの始点に存在す

図 3.3: 商品レビュー投稿画面

るノードが＜主語＞，＜終点＞に存在するノードが目的語となる（図 3.7）．コンテンツデータセットのデータ項目をコンテンツ，クリエイター，ユーザの 3 者関係により説明するため，ノードにはこれら 3 者を設定し，エッジにはこれら 3 者の関係性を設定する．

図 3.4: クックパッドにおけるレシピ説明

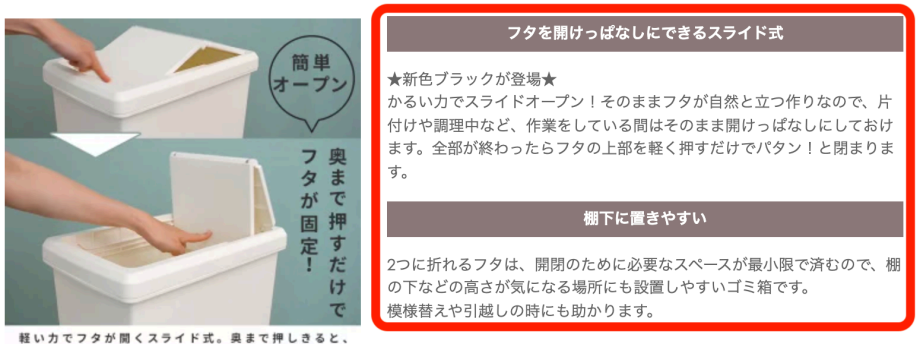

図 3.5: 楽天市場における商品説明

図 3.6: レシピ説明画面

図 3.7: ＜主語＞＜述語＞＜目的語＞の関係

4　対象とするコンテンツデータセットの抽象化

本章では，3 章の提案手法を用い，対象とするコンテンツデータセットを抽象化する．

4.1　データセットの詳細

提案手法により抽象化するコンテンツデータセットは楽天株式会社[1]が提供する楽天市場[2]データセット，楽天トラベル[3]データセット，楽天レシピ[4]データセットの 3 種類に加えて，クックパッド株式会社[5]が提供するクックパッド[6]データセットの計 4 種類とした．これらのデータセットは 4 種類すべてにおいて階層構造となっている[7][8]．例えば，図 4.1 で示すとおり楽天市場データセットは「商品データ」,「みんなのレビュー・口コミ情報」,「市場店舗レビューデータ」の項目から構成されている．「商品データ」の項目内には「商品名」,「店舗コード」,「商品コード」などのデータ項目があり，最下層であるデータ項目名を用いて抽象化を行う．

4.2　コンテンツデータセットの有向グラフ化

3.4 節で提示した RDF 形式を用いて，各コンテンツデータセットのデータ項目を有向グラフ化する．データ項目は，ユーザ，クリエイター，コンテンツ単体に紐付くデータ項目であるか，3 者の関係に紐づくデータ項目であるかの 2 種類に分類した．単体に紐づくデータ項目はノードに配置し，関係に紐づくデータ項目にはエッジに配置する．配置の基準は，各データ項目におけるユーザ，クリエイター，コンテンツの 3 者が担っている役割とする．例えば,「投稿者 ID」のデータ項目は，作成したモノを投稿する役割であるため，クリエイターのノードに配置する．一方で,「レシピの手順」のデータ項目は，クリエイターがコンテンツを説明している項目であるため，クリエイターとコンテンツを結ぶエッジに配置する．エッジとなる述語については，データ項目に対する考察から，目的をメタ的に説明する「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 種類を用意した．

コンテンツ，クリエイター，ユーザの 3 者関係で表すことが可能なデータ項目であっても表現形式によって異なる性質を有している場合がある．例えば，楽天市場の評価ポイントとレビュー内容はどちらもユーザがコンテンツを評価しているデータ項目である．しかし，評価ポイントは離散値であり，レビュー内容はテキストで構成されている．これらの項目は同一の意味的な性質であっても，データ項目の表現形式が異なるため，本稿では別

[1]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: https://doi.org/10.32130/idr.2.0
[2]楽天市場 HP: https://www.rakuten.co.jp（2022/02/14 確認）
[3]楽天トラベル HP: https://travel.rakuten.co.jp（2022/02/14 確認）
[4]楽天レシピ HP: https://recipe.rakuten.co.jp（2022/02/14 確認）
[5]クックパッドデータ, 国立情報学研究所情報学研究データリポジトリ: https://doi.org/10.32130/idr.2.0（2022/02/14 確認）
[6]クックパッド HP: https://cookpad.com（2022/02/14 確認）
[7]Rakkuten Insstitute of Technology 楽天データ公開:https://rit.rakuten.com/data_release_ja（2022/02/14 確認）
[8]国立情報学研究所 情報学データリポジトリ クックパッドデータ提供に関する利用規約 https://cookpad.com/terms/cookpad_data/academy（2022/02/14 確認）

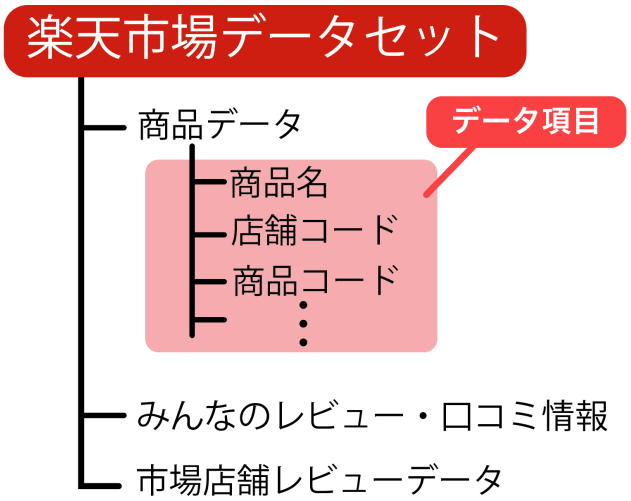

図 4.1: 楽天市場データセットにおけるデータ階層図

項目として考える．

以上の点に留意し，各コンテンツデータセットを RDF 形式の有向グラフで表現したものを図 4.2, 図 4.3, 図 4.4, 図 4.5 に記す．図 4.2，図 4.3，図 4.4 より，ユーザがコンテンツ自体を評価する関係となっているデータ項目は，「レビュー内容」，「評価ポイント」，「おすすめコメント」，「ユーザ投稿本文」などが挙げらる．同様に，図 4.2，図 4.4，図 4.5 より，クリエイターがコンテンツ自体を説明するデータ項目は「商品説明文」，「レシピのきっかけ」，「プランタイトル」，「レシピの生い立ち」などが挙げられる．

有向グラフ化した結果，他のコンテンツデータセットと共通するデータ項目は以下の 7 種類の RDF 形式 （＜主語＞が＜目的語＞を＜述語＞する）に分類された（以下，共通項目と記す）．

共通項目 I　ユーザ　が　コンテンツ　を　evaluate（離散値）する

共通項目 II　ユーザ　が　コンテンツ　を　evaluate（text）する

共通項目III　ユーザ　が　コンテンツ　を　use　する

共通項目IV　クリエイター　が　ユーザ　を　reply　する

共通項目 V　クリエイター　が　コンテンツ　を　describe（離散値）　する

共通項目VI　クリエイター　が　コンテンツ　を　describe(text)　する

共通項目VII　クリエイター　が　コンテンツ　を　create　する

上述の 7 種のデータ項目の分類に該当する，各コンテンツデータセットの各データ項目を表 4.1 に示す．

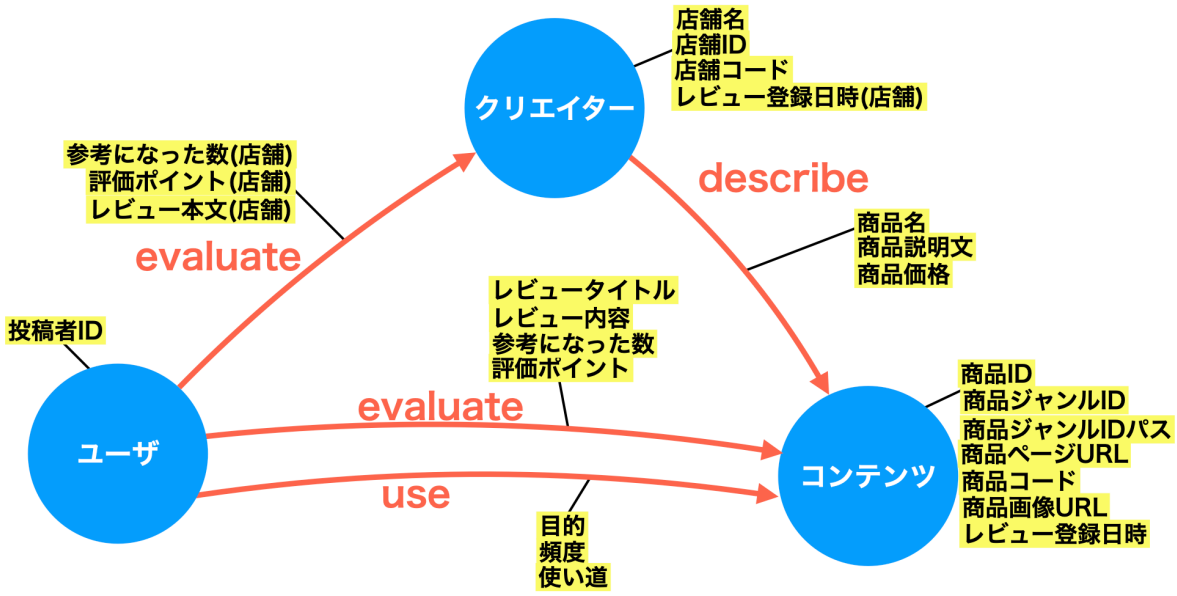

図 4.2: 楽天市場データセットにおけるノードとエッジ図

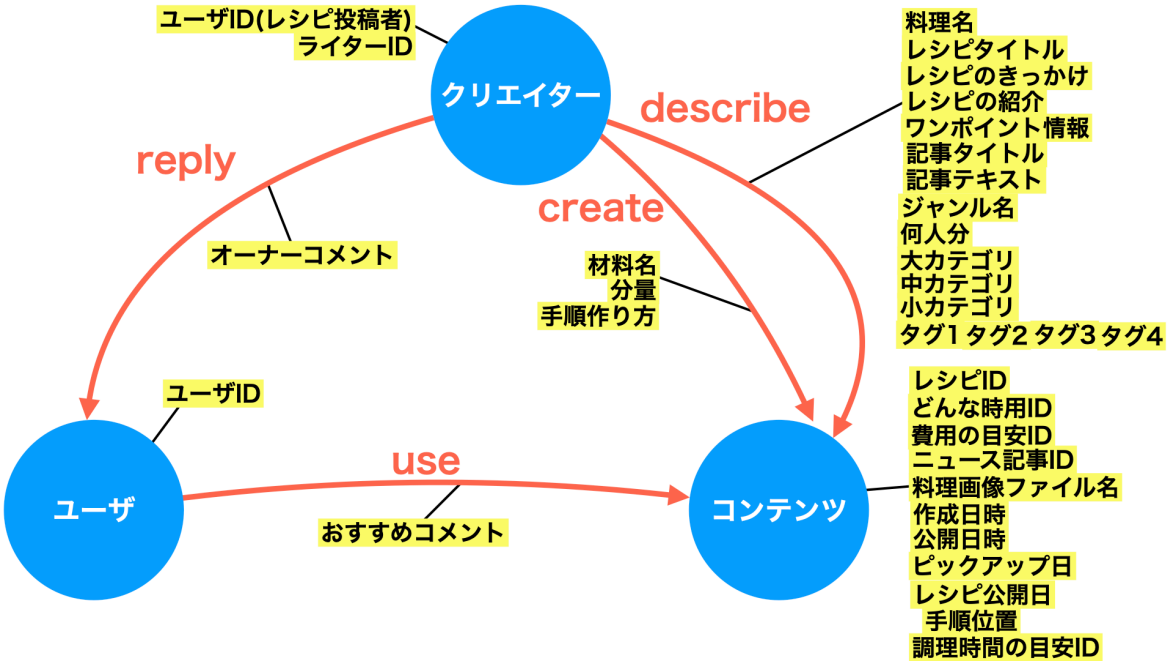

図 4.3: 楽天レシピデータセットにおけるノードとエッジ図

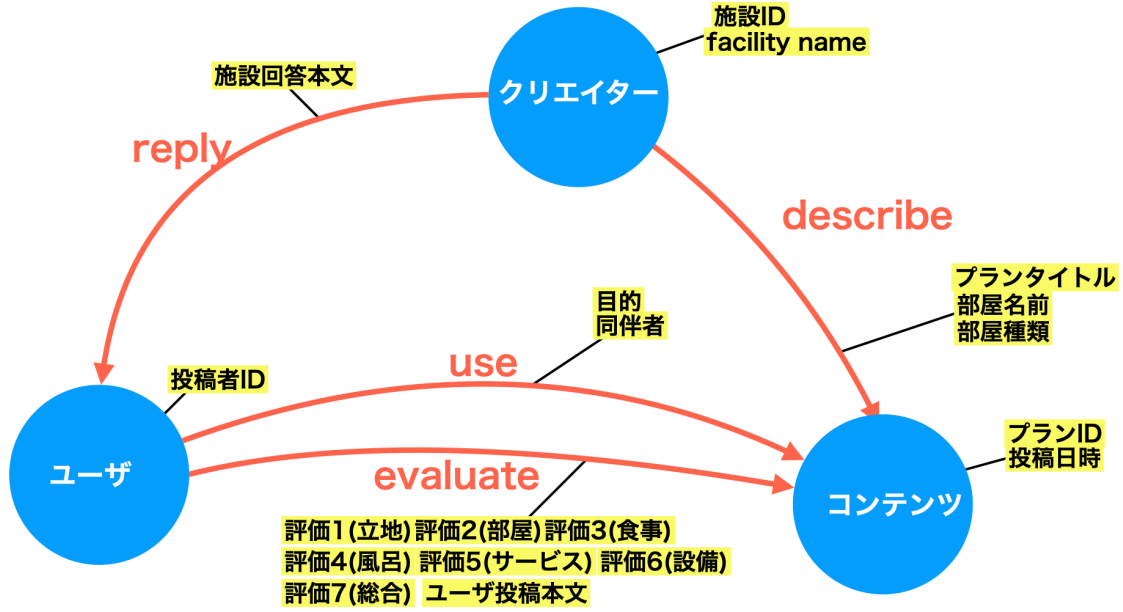

図 4.4: 楽天トラベルデータセットにおけるノードとエッジ図

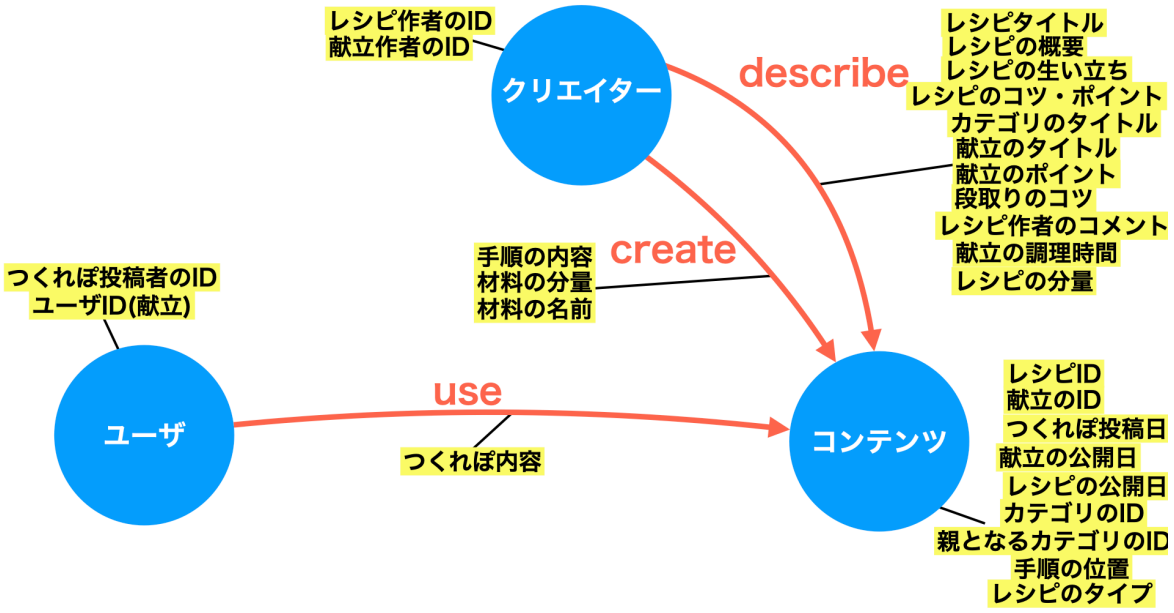

図 4.5: クックパッドデータセットにおけるノードとエッジ図

表 4.1: 各データセットにおける共通データ項目

	主語	目的語	述語	楽天市場	楽天トラベル	楽天レシピ	クックパッド
I	ユーザ	コンテンツ	evaluate(離散)	評価ポイント 参考になった数	評価 1,2,3,4,5,6,7		
II	ユーザ	コンテンツ	evaluate(text)	レビュータイトル レビュー内容	ユーザ投稿本文		
III	ユーザ	コンテンツ	use	目的 頻度 使い道	目的 同伴者	おすすめコメント	つくれぽ内容
IV	クリエイター	ユーザ	reply		施設回答本文	オーナーコメント	
V	クリエイター	コンテンツ	describe(離散)	商品価格		何人分	献立の調理時間 レシピの分量
VI	クリエイター	コンテンツ	describe(text)	商品名 商品説明文	部屋名前 部屋種類 プランタイトル	料理名 レシピタイトル レシピのきっかけ レシピ紹介 ワンポイント情報 記事タイトル 記事テキスト ジャンル名 大，中，小カテゴリ タグ 1, 2, 3, 4	レシピタイトル レシピの概要 レシピの生い立ち レシピのコツ・ポイント カテゴリのタイトル 献立のタイトル 献立のポイント 段取りのコツ レシピ作者のコメント
VII	クリエイター	コンテンツ	create			材料名 分量 手順の作り方	手順の内容 材料の分量 材料の名前

5　提案フレームワークによるデータ項目の分類の妥当性

提案フレームワークにより，分類したコンテンツデータセットのデータ項目が抽象化できているか検証を行う．データ項目の分類の妥当性を示すために，提案フレームワークにより抽象化されたデータ項目を用いて研究を行った論文を使用する．

本検証で使用する論文は，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，クックパッドデータセットの 4 種類のデータセットを用いて研究を行った論文である．これらの 4 種類の論文は，国立情報学研究所の情報学研究データリポジトリから収集した．収集した論文本数は，コンテンツデータセットにつき各 25 本，合計 100 本である．用意した各論文の記述から，研究で利用されたデータ項目を人手で抽出し，その情報を各論文に付与した．

データセットは階層的になっており，複数の属性下にデータ項目が存在する．明確にデータ項目が記載されていない場合は，用いたデータセットの属性下にある全てのデータ項目を用いたと仮定した．

5.1　検証方法

データ項目の抽象化の妥当性を証明するために，ある共通項目を用いた論文を基準としたとき，「その論文と同じ共通項目かつ別のコンテンツデータセットを用いて研究している論文」と類似していると仮定し，検証を行う．検証を行うために，各コンテンツデータセットにおける共通項目を用いた論文を特徴化し，類似性を確認する．

各コンテンツデータセットは扱うコンテンツが異なるため，出現単語を算出した際に，コンテンツに依存した単語が頻出することが考えられる．これにより他のコンテンツデータセットを用いた研究と比較した際，例え共通部となり得る箇所が存在していた場合でも，コンテンツに依存した単語によって埋もれてしまう．また，すべての論文において共通で用いられる単語も出現単語を算出した際に，各論文の特徴となり得ない単語の頻出が想定される．これらの単語は論文そのものの特徴となる単語である．

以上より，異なるコンテンツデータセットを用い，かつ共通項目を用いた論文が類似しているか確認するために，以下のステップを踏む．

（1）コンテンツに依存した単語の除去．

（2）全ての論文で文書頻度数の高い単語の削除．

（3）出現単語を用い類似性を確認．

(1) コンテンツに依存した単語の除去

コンテンツに依存した単語を除去するためにこれらの単語を特定する必要がある．単語を選定するためには，各コンテンツデータセットに依存した単語を見つける必要がある．例えば，料理レシピに関するコンテンツデータセットを用いた論文は，「じゃがいも」や「塩」といった料理や，「容器」や「スプーン」といった料理に関連する道具といった単語が頻出されることが想定される．また，旅行に関するコンテンツデータセットを用いた論文は，「ト

イレ」や「温泉」といった施設の設備に関する単語が頻出されることが想定される．論文の頻出単語を除去した場合，コンテンツに関連しない頻出単語も除去する対象となる可能性がある．的確にコンテンツに依存した単語を特定するために単語を意味的に分類し，同様の意味で単語の集合（以下，単語群とする）を作成する．コンテンツに依存したと判断される単語群を選択し，その単語群内の単語すべてを単語を特定する．コンテンツに依存した単語の特定プロセスを図 5.1 に記載する．100 本すべての論文内に含まれる単語を意味的に分類し単語群を作成する（図 5.1-①）．得られた単語群を辞書として単語の出現頻度から 100 本すべての論文を特徴化する．特徴化した論文を各コンテンツデータセットごとでまとめ，コンテンツデータセットを用いた論文 25 本ずつで，どの単語群が出現したか割合を算出する（図 5.1-②）．コンテンツデータセットごとに算出した単語群の出現割合から分散を算出する（図 5.1-③）．分散の値が大きい単語群は，特定のコンテンツデータセットにしか単語群内の単語が出現していないということを表している．本稿では，それらの単語群内の単語をコンテンツに依存した単語とする．これらのことから，コンテンツに依存した単語を特定するために以下のステップを踏むこととする．

(1)-1 すべての論文の本文情報に対して単語を意味的に分類し，単語群を作成．

(1)-2 各コンテンツデータセットを用いた論文において，どの単語群が頻出しているか算出．

(1)-3 各コンテンツデータセットごとに単語群の分散を算出し，分散が大きい単語群を確認．

(1)-1 すべての論文の本文情報に対して単語を意味的に分類し，単語群を作成．

論文に出現する単語を意味的に分類し，単語群を作成するためにクラスタリングを行う．

まず，100 本すべての論文の本文情報を形態素解析器 MeCab(ver. 0.996)[1]を用いて単語分割を行った．その際，論文には固有表現が多数含まれているため，辞書には固有表現に強い mecab-ipadic-neologd[2]辞書を用いた． また，指示語といった，論文自体の意味を示さない単語が含まれることを防ぐために，SlothLib[19] に含まれる単語をストップワードとした．

クラスタリングを行う上で，全ての論文内でほとんど使用されていない固有名詞や単語などはクラスタリングを行う上でノイズとなるため除去する．それらの単語を特定するために，文書頻度数を表す DF 値を算出した．DF 値の算出方法は式 (5.1) の通りである．

$$DF_i = \frac{df_i}{N} \tag{5.1}$$

N は総文章を示している．df_i は単語 i を含む文書数である．DF 値を降順に並べ確認したところ論文の単語全体の 50%はロングテール状態となっていた（図 5.3 参照）．DF 値を確認したところ，0.01 であったため本稿では DF 値が 0.01 より高い単語を用いて分類する．

[1] https://taku910.github.io/mecab/（2022/1/6 存在確認）
[2] https://github.com/neologd/mecab-ipadic-neologd（2022/1/6 存在確認）

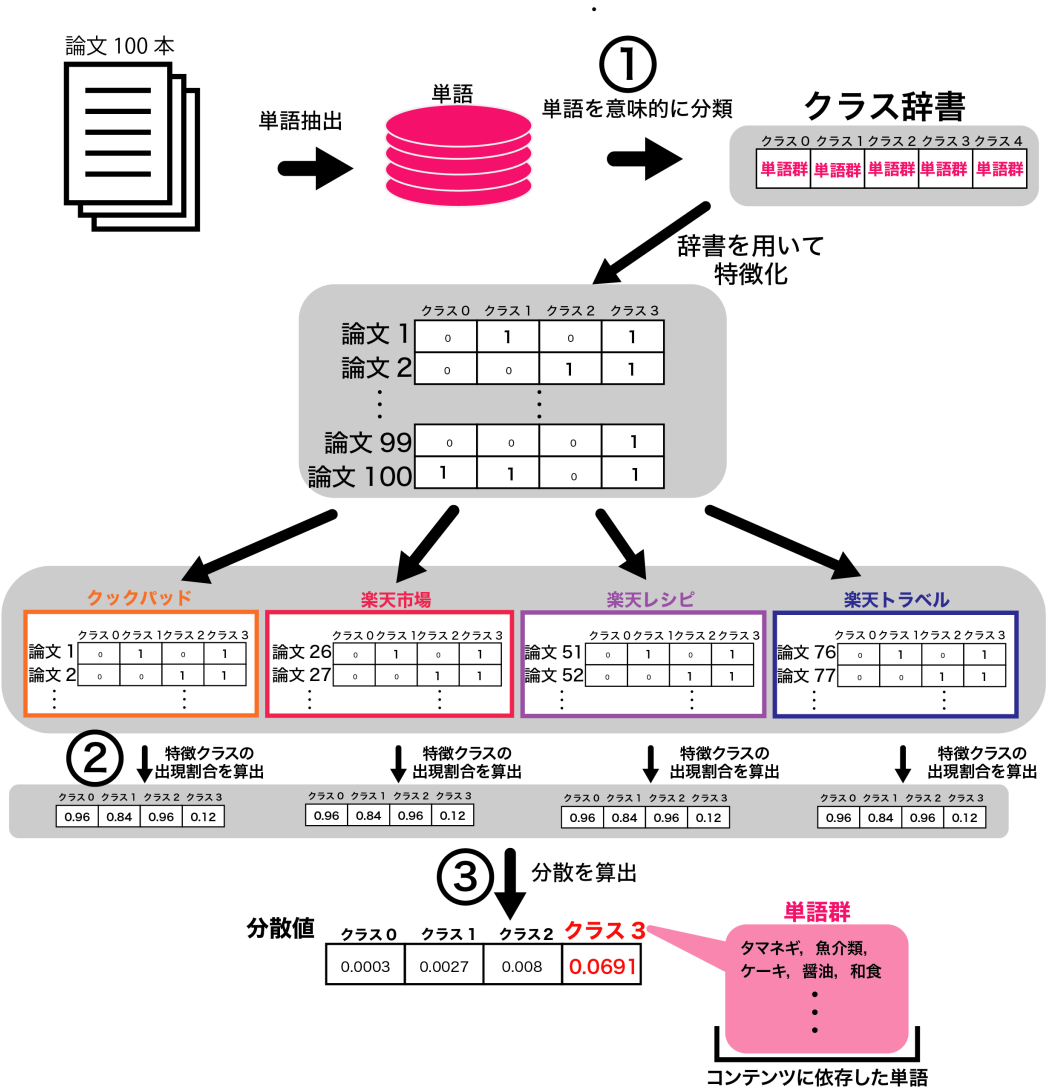

図 5.1: コンテンツに依存した単語の特定の流れ

次に，単語を意味的に分類する．用意したデータセットの単語をベクトル化するために，本稿では鈴木らが作成した日本語 Wikipedia エンティティベクトル [23] を用いることとする．この Wikipedia ベクトルは日本語版 Wikipedia の本文を学習データとして構築されている．

得られた単語ベクトルから，k-means++法 [1] を用いて単語をクラスタリングした．k-means++法を用いた理由として，通常の k-means 法 [7] では初期値をランダムに設定してしまうという欠点がある．k-means++法は，クラスタ中心の初期値を決定する際に，クラスタの中心点同士が遠いところに配置される確率が高いアルゴリズムを採用している [25]．これらの理由から，ランダムに初期値を選択する k-means 法よりもより良いクラスタリングが実現可能な k-means++法を用いることとする [21]．k-means++法のクラス数はエルボー法 [3] 用いて選定をしたところ，45 クラスという結果が得られた．

(1)-2 各コンテンツデータセットを用いた論文において，どの単語群が頻出しているか算出．

各論文の内容を把握するために，作成した単語群を用い，論文内の単語の分類を行う．単語群内の単語の出現の有無で各論文ごとにバイナリ列を作成する．バイナリ列は，単語群内の単語が出現した場合に 1 を付与し，出現していない場合は 0 を付与することで作成を行う．低頻出である単語にまで 1 を付与することは，その論文において特徴的ではない単語にまで意味づけを行うことになるため，特徴的ではない単語を基準値を作成する．頻度をもとにした基準値を作成する．特徴的な単語を選定するための基準値を設定するために，今回は TF-IDF 法を用いることとする [4]．TF-IDF 法とは統計的な情報検索手法において文献のタイトル，抄録，本文などを語単位に分割し，各語の重みを計算する方法のことである [26]．TF-IDF 値は単語の出現頻度を表す TF 値と逆文書頻度数を表す IDF 値の積から算出する．

TF 値の算出方法は式 (5.2) の通りである．

$$TF_{ij} = \frac{n_{ij}}{X} \tag{5.2}$$

$n_{i,j}$ は文章 j における単語 i を示している．X は文章 j に出現する単語の総頻出度である．

IDF 値の算出方法は式 (5.3) の通りである．

$$IDF_i = \log_e \frac{N}{df_i} + 1 \tag{5.3}$$

N は総文章を示している．df_i は単語 i を含む文書数である．

式 (5.2) と式 (5.3) の積より TF-IDF が算出可能である（式 (5.4) 参照）．

$$TFIDF_{ij} = TF_{ij} \cdot IDF_i \tag{5.4}$$

各論文に出現した単語の TF-IDF 値から中央値を算出した．TF-IDF 値の中央値を基準とし，その値以上の単語が出現した際に特徴語として単語群に 1 を付与し，それ以外の単語群には 0 を付与することでバイナリ列を作成した．

(1)-3 各コンテンツデータセットごとに単語群の分散を算出し，分散が大きい単語群を確認．

コンテンツに依存した単語を特定するため，作成したバイナリ列を各コンテンツデータセットで特性を確認した．各コンテンツデータセットを用いた論文 25 件でバイナリ列の割合を算出した．各コンテンツデータセットの平均したバイナリ列の項目で分散を算出し，どの単語群がコンテンツに依存しているか確認した．分散を算出したものを度数分布表にし示す (図 5.2 参照)．横軸は分散値，縦軸は単語群の数を表している．図 5.2 より，分散の値が 0.000 以上から 0.004 未満に集中していることが確認できる．分散の値が 0.004 以上の単

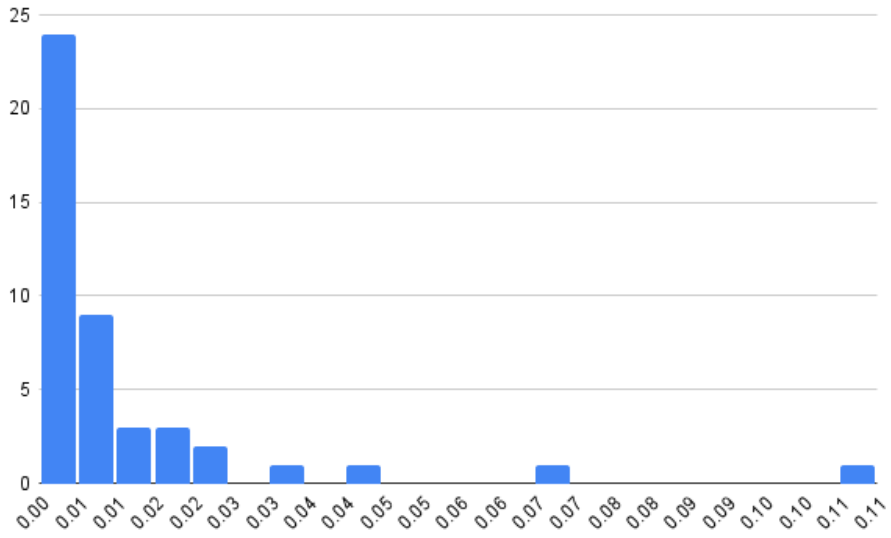

図 5.2: 各クラスにおける分散

語群の分散の値が大きい傾向にあるため，これらの単語群内に含まれる単語を除去することとする．0.004 以上の単語群は 3 つ存在し，「ホテル」や「宿泊施設」といった施設に関する単語群，「購買」や「小売店」といったショッピングに関するクラス，「タマネギ」や「肉」といった料理に関する単語群が確認できた（表 5.1 参照）．料理に関する単語群ではクックパッドと楽天レシピのデータセットを用いた論文のすべてに 1 が付与されていた．ショッピングに関するクラスでは，楽天市場のデータセットを用いた論文の 92.0%に 1 が付与されていた．施設に関するクラスでは，楽天トラベルのデータセットを用いた論文の 88.0%に 1 が付与されていた．4 つのコンテンツに依存する単語をこれら 3 つのクラス内に含まれる単語とする．

(2) 全ての論文で文書頻度数の高い単語の削除

全ての論文内で出現する単語を取り除くため，文書頻度数を表す DF 値を用い確認する（図 5.3 参照）．DF 値が高い単語は，すべての論文で頻出するものであり，本稿では DF 値が 0.7 を基準とし，その値以上の単語を除去することとした．

(3) 出現単語を用い類似性を確認

類似性を確認するため，検証で用いる 100 本の論文の中から共通項目が同じで別のコンテンツデータセットを用いた論文ペアを確認する．今回は，以下の 4 ペア 8 種別の論文を用いて検証を行う．

- 共通項目Ⅵ「クリエイターがコンテンツを describe(text) する」と共通項目Ⅶ「クリエイターがコンテンツを create」の両方を扱ったクックパッドデータセットを用いた論文と楽天レシピデータセットを用いた論文
- 共通項目Ⅶ「クリエイターがコンテンツを create」のみを扱ったクックパッドデータ

表 5.1: 分散の値が 0.004 以上の単語群内の単語の一例

単語群 A	単語群 B	単語群 C
ホテル	製品	塩
トイレ	値段	和食
宿泊	取引	タレ
地域	売上	タマネギ
観光地	価格	肉
駅	売買	魚介類
温泉	小売店	揚げ物
観光名所	購買	お菓子
施設	顧客	薬味
レジャー	商品	アイス

セットを用いた論文と楽天レシピデータセットを用いた論文

- 共通項目 I「ユーザがコンテンツを evaluate(離散)」と共通項目 II「ユーザがコンテンツを evaluate(text)」の両方を扱った楽天市場データセットの論文と楽天トラベルの論文
- 共通項目 II「ユーザがコンテンツを evaluate(text)」のみを扱った楽天市場データセットの論文と楽天トラベルデータセットの論文

これらの検証には 8 種別各 5 論文，合計 40 本の論文を用いることとする．共通項目を用いた論文がどの論文と類似しているか検証する．そのために，各コンテンツデータセットの 5 本の論文の名詞のみの単語から TF 値を算出し，その平均値を取る．TF 値の度数分布表を 8 種別で作成する．その度数分布表から総当りで類似度を算出する．

度数分布表の類似性を表す事が可能である Bhattacharyya 距離を算出する [2][22]．計算式はコンピュータビジョン向けライブラリの Open CV2.2 内にある calcHist 関数を使用した[3]（式 5.5 参照）．提案手法で算出した識別特徴語の章における分布と，選択語の章における分布から Bhattacharyya 距離を算出した．

$$d(H_1, H_2) = \sqrt{1 - \frac{1}{\sqrt{\overline{H}_1 \overline{H}_2 N^2}} \sum_I \sqrt{H_1(I) \cdot H_2(I)}}, \tag{5.5}$$

ここで，N は区間の数を表し，H_1，H_2 は各区間の値を，$\overline{H}_1$，$\overline{H}_2$ は平均値を表している．

5.2 検証結果

算出した Bhattacharyya 距離の結果を表 5.2 に示す．表 5.2 より，行ラベルに記載してある共通項目を用いた論文を基準とした際の最も Bhattacharyya 距離が近かった数値に下線

[3] http://opencv.jp/opencv-2svn/cpp/histograms.html（2022/1/6 存在確認）

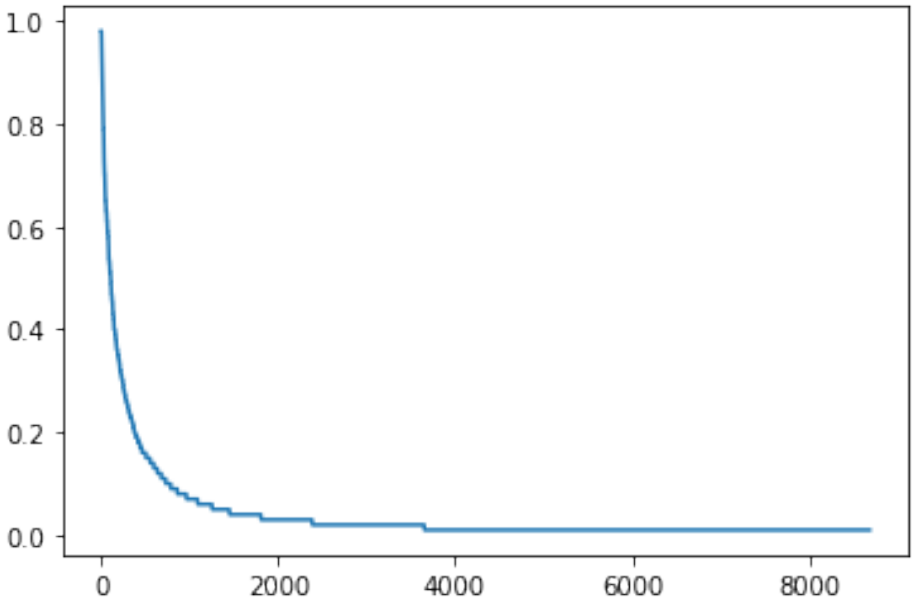

図 5.3: 100 本の論文における DF 値

を引き示す．ここで，1 つの論文を基準とし同じ共通項目を用いた別のコンテンツデータセットを用いた論文との類似距離の近さの順位を確認する．

共通項目Ⅵ「クリエイターがコンテンツを describe(text) する」と共通項目Ⅶ「クリエイターがコンテンツを create する」の両方を扱ったクックパッドの論文を基準としたとき，共通項目Ⅵ「クリエイターがコンテンツを describe(text) する」と共通項目Ⅶ「クリエイターがコンテンツを create する」を扱った楽天レシピの論文との類似距離の順位は 2 番目に近くなった．また，共通項目Ⅵ「クリエイターがコンテンツを describe(text) する」と共通項目Ⅶ「クリエイターがコンテンツを create する」の両方を扱った楽天レシピの論文を基準としたとき，共通項目Ⅵ「クリエイターがコンテンツを describe(text) する」とⅦの項目を持ったクックパッドの論文との類似距離は 2 番目に近くなった．

共通項目Ⅶ「クリエイターがコンテンツを create する」のみを扱ったクックパッドの論文を基準としたとき，共通項目Ⅶ「クリエイターがコンテンツを create する」のみを扱った楽天レシピの論文との類似距離は 3 番目に近くなった．また，共通項目Ⅶ「クリエイターがコンテンツを create する」のみを扱った楽天レシピの論文を基準としたとき，共通項目Ⅶ「クリエイターがコンテンツを create する」のみを扱ったクックパッドの論文との類似距離は最も近くなった．

共通項目Ⅰ「ユーザがコンテンツを evaluate(離散)」と共通項目Ⅱ「ユーザがコンテンツを evaluate(text)」の両方を扱った楽天市場の論文を基準としたとき，共通項目Ⅰ「ユーザがコンテンツを evaluate(離散)」と共通項目Ⅱ「ユーザがコンテンツを evaluate(text)」の両方を扱った楽天トラベルの論文との類似距離は 2 番目に近くなった．また，共通項目Ⅰ「ユーザがコンテンツを evaluate(離散)」と共通項目Ⅱ「ユーザがコンテンツを evaluate(text)」の両方を扱った楽天トラベルの論文を基準としたとき，共通項目Ⅰ「ユーザがコンテンツを evaluate(離散)」と共通項目Ⅱ「ユーザがコンテンツを evaluate(text)」の両方を扱った楽天

市場の論文との類似距離は 3 番目に近くなった．

共通項目 II「ユーザがコンテンツを evaluate(text)」のみ扱った楽天市場の論文を基準としたとき，共通項目 II「ユーザがコンテンツを evaluate(text)」のみ扱った楽天トラベルの論文との類似距離は最も近くなった．また，共通項目 II「ユーザがコンテンツを evaluate(text)」のみ扱った楽天トラベルの論文を基準としたとき，共通項目 II「ユーザがコンテンツを evaluate(text)」のみ扱った楽天市場の論文との類似距離は最も近くなった．

同じ共通項目かつ別のコンテンツデータセットを用いた論文の類似距離は 3 番目以内であり，順位は平均 1.88 位であった．これらの結果から，定量的な検証では提案フレームワークの妥当性が示された．

表 5.2: Bhattacharyya 距離の結果

	Ⅵ&Ⅶ（クックパッド）	Ⅵ&Ⅶ（楽天レシピ）	Ⅶ（クックパッド）	Ⅶ（楽天レシピ）	Ⅰ&Ⅱ（楽天市場）	Ⅰ&Ⅱ（楽天トラベル）	Ⅱ（楽天市場）	Ⅱ（楽天トラベル）
Ⅵ&Ⅶ（クックパッド）	0.000	0.690	0.681	0.714	0.805	0.785	0.735	0.772
Ⅵ&Ⅶ（楽天レシピ）	0.690	0.000	0.682	0.705	0.808	0.779	0.717	0.773
Ⅶ（クックパッド）	0.681	0.682	0.000	0.707	0.810	0.793	0.740	0.779
Ⅶ（楽天レシピ）	0.714	0.715	0.707	0.000	0.797	0.777	0.750	0.793
Ⅰ&Ⅱ（楽天市場）	0.805	0.808	0.810	0.797	0.000	0.760	0.722	0.779
Ⅰ&Ⅱ（楽天トラベル）	0.785	0.779	0.792	0.777	0.769	0.000	0.746	0.759
Ⅱ（楽天市場）	0.735	0.717	0.740	0.750	0.722	0.746	0.000	0.714
Ⅱ（楽天トラベル）	0.772	0.773	0.779	0.793	0.779	0.759	0.714	0.000

6 議論

本研究では，複数のコンテンツデータセットに含まれるデータ項目群を横断的に理解するための情報整理フレームワークをユーザ，コンテンツ，クリエイターの 3 者関係により表現した．本章ではこの 3 者関係に紐づく共通項目の実データを定性的に確認することで，提案フレームワークが妥当であったか議論する

6.1 横断的に関係を把握可能な共通項目

本フレームワークにより横断的に関係を把握可能と判断された共通項目は，以下の 3 項目であった．

Ⅱ「ユーザがコンテンツを evaluate(テキスト) する」

Ⅳ「クリエイターがユーザを reply する」

Ⅶ「クリエイターがコンテンツを create する」

共通項目Ⅱ「ユーザがコンテンツを evaluate(text) する」は，楽天市場データセット，楽天トラベルデータセットで確認された．楽天市場データセットの「レビュー内容」と楽天トラベルの「ユーザ投稿文」の 2 つのデータ項目は共に「ユーザがコンテンツを evaluate(text) する」に該当するためこの 2 つのデータ項目を確認する．楽天市場データセットの「レビュー内容」の実データを例にあげる[1]．

> 1 日に何度も掃除機をかけますがその度にこの軽さに感動しています．

この楽天市場データセットの「レビュー内容」は，「ユーザが＜掃除機＞（コンテンツ）を＜軽い＞ (evaluate)」と評価している．楽天トラベルの「ユーザ投稿文」にのデータ内の例をあげる[2]．

> 部屋が綺麗で良かったです

ユーザ投稿本文は，「ユーザが＜部屋＞（コンテンツ）を＜綺麗＞ (evaluate)」と評価している．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「ユーザ投稿文」は，それぞれのコンテンツに対して「ユーザがコンテンツを evaluate する」の関係にあるといえる．

つまり，この共通項目Ⅱは異なるデータセット間のデータ項目を同一のものとして扱うことが可能であり，異なるデータセット間においても横断的に把握することが可能であった．

共通項目Ⅳ「クリエイターがユーザを reply する」は楽天レシピデータセット，楽天トラベルデータセットで確認された．楽天レシピの「オーナーコメント」と楽天トラベルデー

[1]楽天市場 東芝サイクロン掃除機: https://review.rakuten.co.jp/item/1/243088_10729232/1.1/(2022/02/12 確認)

[2]楽天トラベル 神戸メリケンパークオリエンタルホテル: https://travel.rakuten.co.jp/HOTEL/8978/review.html(2022/02/12 確認)

タセットの「施設回答本文」は,「ユーザがコンテンツを evaluate(text) する」に該当するため，この 2 つのデータ項目を確認する．

楽天レシピデータセットの「オーナーコメント」の実データを例にあげる[3]．

> 簡単なのでぜひまた作ってください

「また作って」という表現から，ユーザがレシピ（コンテンツ）を作成したことに対し，クリエイターがユーザに返答 (reply) していると言える．楽天トラベルデータセットの「施設回答本文」の実データを例にあげる[4]．

> またのご来館をスタッフ一同お待ち申し上げております。

「またの」や「お待ち申し上げております」という表現から，ユーザが施設（コンテンツ）を利用したことに対しクリエイターがユーザに返答 (reply) していると言える．楽天レシピデータセットの「オーナーコメント」と，楽天トラベルデータセットの「施設回答本文」は，「クリエイター」＜主語＞が「ユーザが関わったこと」＜目的語＞に「reply」＜述語＞していたと言える．つまり，この共通項目Ⅳは異なるデータセット間のデータ項目を同一のものとして扱うことが可能であり，異なるデータセット間のデータ項目を横断的に把握することが可能であった．

共通項目Ⅶ「クリエイターがコンテンツを create する」は楽天レシピデータセット，クックパッドデータセット全てのコンテンツデータセットで確認された．楽天レシピでは「材料名」「手順の作り方」といった項目，クックパッドでは「材料の名前」「手順の内容」といった項目で構成されている．楽天レシピデータセットとクックパッドデータセット内で同一の項目が存在しており，その内容にも相違ないことから提案フレームワークにより分類ができていると言える．

つまり，この共通項目Ⅶは異なるデータセット間においても横断的に把握することが可能であった．

提案フレームワークでは共通項目Ⅱ「ユーザがコンテンツを evaluate(text) する」と共通項目Ⅳ「クリエイターがユーザに reply する」と共通項目Ⅶ「クリエイターがユーザに create する」に関して，横断的に関係を把握可能であることが示唆された．

6.2 一部横断的に関係を把握可能な共通項目

本フレームワークにより一部横断的に関係を把握可能と判断された共通項目は，以下の 4 項目であった．

Ⅰ「ユーザがコンテンツを evaluate(離散) する」

Ⅲ「ユーザがコンテンツを use する」

[3]楽天レシピ:10 分でできる！ うちのしょうが焼きつくったよ:https://recipe.rakuten.co.jp/recipe/1510010731/report/(2022/02/12 確認)

[4]楽天トラベル ホテルグランヴィア京都:https://travel.rakuten.co.jp/HOTEL/1487/review.html(2022/02/12 確認)

Ⅴ「ユーザがコンテンツを describe（離散）する」

Ⅵ「クリエイターがコンテンツを describe（text）する」

データセットにおける共通項目Ⅰ「ユーザがコンテンツを evaluate(離散) する」は楽天市場データセット，楽天トラベルデータセットで確認された．楽天市場データセットの「評価ポイント」と楽天トラベルの「評価 1 ～ 7」の 2 つのデータ項目は共に「ユーザがコンテンツを evaluate(離散) する」に該当するためこの 2 つのデータ項目の実データを確認する．楽天データセットの「評価ポイント」と楽天トラベルの「評価 1 ～ 7」は共にユーザがコンテンツに対し評価したものが離散値の評価として現れているため，これらの項目は「ユーザがコンテンツを evaluate(離散) する」の関係にあるといえ，本フレームワークにより横断的に把握可能である．一方で，共通データ項目Ⅰ「ユーザがコンテンツを evaluate(離散) する」の共通項目である楽天市場データセット内の「参考になった数」のデータ項目は「評価ポイント」と「評価 1 ～ 7」と同様に離散値という性質を持つが，評価しているコンテンツが異っていた．例えば，ユーザが購入した＜掃除機＞に対してレビュー文が記載されていた場合,「参考になった数」のデータ項目は＜掃除機＞というコンテンツに対しての評価値ではなく,「掃除機を評価したレビュー」に対しての評価値である．

つまり,「参考になった数」のデータ項目は「＜ユーザ＞が＜コンテンツ＞を evaluate」したものに対し，さらに evaluate(離散）している項目となっている．これは,「レビュータイトル」や「レビュー内容」といったデータ項目に対してユーザが evaluate していると言える．楽天市場データセット内の「参考になった数」は「＜主語＞は＜目的語＞を＜述語＞することを＜述語＞する」の関係となり，いずれの共通項目にも該当しない．

データセットにおける共通データ項目Ⅲ「ユーザがコンテンツを use する」は楽天市場データセット，楽天トラベルデータセットで確認された．クックパッドデータセットの「つくれぽ内容」と楽天レシピデータセットの「おすすめコメント」は共に．「ユーザがコンテンツを use する」に該当するため，この 2 つのデータ項目を確認する．クックパッドデータセットの「つくれぽ内容」[5]，の実データを例にあげる．

> ワカモレ美味しく出来ました！

同様に，楽天レシピデータセットの「おすすめコメント」[6]．の実データを例にあげる．

> 簡単にできました！

「つくれぽ内容」と「おすすめコメント」のデータ項目にはコンテンツを使用した感想が記述されていた．よって,「つくれぽ内容」と「おすすめコメント」のデータ項目においては横断的に関係性を読み解くことができる．

共通項目Ⅲ「ユーザがコンテンツを use する」のデータ項目である楽天市場データセットの「使い道」のデータ項目には【趣味】や【イベント】といった内容が記載されており，

[5]クックパッド 簡単ワカモレで メキシコの本格タコス: https://cookpad.com/recipe/3931567（2022/02/14 確認）

[6]楽天レシピ 豚肉の生姜焼き レシピ・作り方: https://recipe.rakuten.co.jp/recipe/1020000132/report/4/(2022/02/12 確認)

楽天トラベルの「同伴者」の項目には【一人】や【家族】といった内容が記載されている．これらの項目は単語であり，単語単体でユーザがコンテンツを用いたかどうかは把握することはできない．共通項目Ⅲは一部のコンテンツデータセットのデータ項目を横断的に把握することが可能であったが，一部のデータ項目では横断的に把握することは困難である．

共通項目Ⅴ「クリエイターがコンテンツを describe(離散) する」は楽天市場データセット，楽天レシピデータセット，クックパッドデータセットで確認された．楽天レシピデータセットの「何人分」とクックパッドデータセットの「レシピの分量」の 2 つのデータ項目は共に「クリエイターがコンテンツを describe(離散) する」に該当し，共にコンテンツが何人用の分量かを離散値で記載されている．そのため，「クリエイターがコンテンツを describe する」という項目において，これらの関係性は横断的に把握可能である．一方で，楽天市場データセットの「商品価格」とクックパッドの「献立の調理時間」はどちらも離散値データであるが，価格 (e.g, 円) や時間 (e.g, 分) といった単位が異なるため，これらの関係性は読み解くことできない．ただし，「クリエイターがコンテンツを describe(離散) する」は離散値の単位が同じデータ項目でであれば，横断的に把握可能であると言える．

共通項目Ⅵ「クリエイターがコンテンツを describe(text) する」は楽天市場データセット，楽天レシピデータセット，楽天トラベルデータセット，クックパッドデータセットの 4 つのコンテンツデータセットで確認された．

クックパッドデータセット内の「レシピタイトル」と楽天トラベル内の「プランタイトル」と「クリエイターがコンテンツを describe(text) する」は，共に一言で各コンテンツの説明を行っている．同様に，楽天市場の「商品説明文」のデータ項目とクックパッドの「レシピの生い立ち」の 2 つのデータ項目を確認する．「商品説明文」の実データを例にあげる．

> 強力な気流でゴミを圧縮（トルネードプレス）しネット部へのゴミの付着も低減します。

この「商品説明文」は，「クリエイターがコンテンツに対しゴミの付着も低減できる」と説明しているしている．「レシピの概要」[7]．の実データを例にあげる．

> 家でなかなかふかふかのホットケーキは焼けなかったけど、ホットサンドメーカーで焼いたら分厚いホットケーキが焼けました

この「レシピの概要」は，「クリエイターがホットケーキ（コンテンツ）に対しホットケーキが焼けた」と説明している．これらの「商品説明文」と「レシピの概要」は共にコンテンツに対しクリエイターが説明を行っていると言える．一方で，共通項目Ⅵ「クリエイターがコンテンツを describe(text) する」の楽天市場データセットの「料理名」には，【和風パスタ】といった内容が記載してある．これらの項目は単語であり，単語単体でクリエイターがコンテンツを説明したかは把握することができなかった．

以上の定性的な調査から，共通項目Ⅰ〜Ⅶに分類されたデータ項目の実データは，それぞれに対応した「＜主語＞が＜目的語＞を＜述語＞する」の関係であるデータ項目が存在した．そのため，これらの共通項目においてコンテンツデータセット間の関係性を横断的に把握することが可能であると言える．ただし，共通データ項目Ⅰ, Ⅲ, Ⅴ, Ⅵに分類された

[7] クックパッド ふわふわホットケーキ: https://cookpad.com/recipe/7099814（2022/02/12 確認）

データ項目の実データを確認したところ，一部のデータ項目はどの共通項目にも該当しなかったことから，コンテンツデータセット間の一部のデータ項目においては関係性を横断的に把握することはできなかった.

7　結論

本稿では，複数のコンテンツデータセットに含まれるデータ項目群を横断的に理解するための情報整理のフレームワークを提案した．コンテンツデータセットに含まれるデータ項目の性質を，コンテンツ自身，コンテンツを提供するクリエイター，コンテンツを利用するユーザの 3 者の関係によって抽象化を行った．妥当性の検証として，ある共通項目を用いた論文は，その論文と同じ共通項目かつ別のコンテンツデータセットを用いて研究している論文と類似してるという仮定のもと定量的な検証を行い，実データの中身を確認することで定性的な検証を行った．ある共通項目を用いた論文は，同一の共通項目かつ別のコンテンツデータセットを用いた論文の Bhattacharyya 距離の平均順位は平均 1.88 位となり，定量的な検証では提案フレームワークの妥当性が示された．また，定性的な検証をするために，各共通項目の実データを参照した結果，共通項目Ⅱ，Ⅳ，Ⅶにおいて異なるコンテンツデータセット間を横断的に把握することが可能となり，共通項目Ⅰ，Ⅲ，Ⅴ，Ⅵにおいて異なるコンテンツデータセット間を一部横断的に把握することが可能となった．

本稿の提案フレームワークでは，離散値データとテキストデータは異なるデータ性質を持つため，「ユーザがコンテンツを evaluate」する項目は，それぞれ別項目として扱う必要があった．しかし，提案フレームワークは他の手法と組み合わせて使用することで，＜主語＞＜目的語＞＜述語＞が同一であれば異なるデータ性質を同一の共通項目として扱える可能性がある．例えば，レビュー文から「綺麗」や「便利」といった評価に関するキーワードから評価表現辞書を用いて自動で評価値として算出する，というように他の技術を用いることでレビュー文を離散値での表現が可能になる．テキストデータを離散値データでとして扱うことが可能であれば，データの表現形式が異なる場合でも共通項目として扱うことが可能となる．

本稿で提案した情報整理のためのフレームワークでは，コンテンツデータセットのデータ項目 1 件につき，1 つの共通項目を付与する事となった．しかし，クックパッドデータセットの「つくれぽ内容」といったデータ項目は共通項目Ⅲ「ユーザがコンテンツを use する」とⅥ「ユーザがコンテンツを evaluate(text) する」の両方に該当する可能性がある．本稿では複数の共通項目に該当する場合を考慮しなかったため，今後は，複数の共通項目に該当するデータ項目を抽出し検証していく必要がある．

謝辞

この研究は 2021 年度国立情報学研究所公募型共同研究 (21S0501) の助成を受けました．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/) を利用させていただきました．本研究では、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クックパッド株式会社から提供を受けた「クックパッドデータセット」を利用させていただきました．

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．並びに，関西大学総合情報学部の山西良典准教授にも数々のご指導ご鞭撻を賜りました．松下教授には研究活動はもちろんのこと，社会で必要となる知識や心持ちを教えていただきました．心より深く感謝致します．

関西大学大学院総合情報学研究科知識情報学専攻の小林光氏，福元颯氏，森野穣氏，樋口友梨穂氏，宮本誠人氏，樋口亮太氏，竹村孟氏，山本京香氏に心より感謝申し上げます．小林光氏は，AO 入試からお世話になりました．小林氏は大学で出来た友達の中で最も長い付き合いになりました．福元颯氏は，プログラミング入門の講義からお世話になりました．東京から京都まで自転車で旅をした際，励ましていただいたのを今でも覚えております．森野穣氏は，研究室配属から大変お世話になりました．一緒に釣りに行ったり映画を見に行ったりとプライベートから研究活動で大変お世話になりました．特に研究活動では森野氏に一番お世話になりました．心より深く感謝致します．樋口友梨穂氏には，日々のメンタルケアに付き合っていただきました．宮本誠人氏には，日々の雑談にお付き合いただきました．心より感謝申し上げます．樋口亮太氏，竹村孟氏，山本京香氏には，日々の研究室での雑談にお付き合いいただいております．

研究室生活を送る上でお世話になりました 8 期生，9 期生，10 期生，11 期生，12 期生の皆様に深く感謝申し上げます．特に，岩崎有基氏，中西聖氏，青木靖太氏，返町周氏，建田伸氏には様々な点でお世話になりました．岩崎氏，中西氏には研究室での立ち振舞方についてご指導いただきました．青木氏には，卒業後も高槻近辺で立ち話に付き合っていただきました．返町氏には論文添削をしていただきました．皆様に深く感謝申し上げます．

太平洋を跨いだ北アメリカ大陸にいる椎原理央氏には小学 1 年生から，お世話になっております．昼夜逆転した際は時差の関係で椎原氏しか起きておらず，いつも雑談に付き合っていただきました．心より深く感謝致します．

最後に，6 年間長い期間の学生生活を遠方から，支えてくれた両親，姉に心から感謝の意を記して謝辞と致します．

参考文献

[1] Arthur, D. and Vassilvitskii, S.: k-means++: The advantages of careful seeding, *Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithm*, pp. 1027–1035 (2007).

[2] Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions, *Bulletin of the Calcutta Mathematical Society*, Vol. 35, pp. 99–109 (1943).

[3] Bholowalia, P. and Kumar, A.: EBK-means: A Clustering Technique based on Elbow Method and K-Means in WSN, *International Journal of Computer Applications*, Vol. 105, No. 9 (2014).

[4] Gerard, S. and Christopher, B.: Term-weighting approaches in automatic text retrieval, *Information processing & management*, Vol. 24, No. 5, pp. 513–523 (1988).

[5] Ghavimi, B., Mayr, P., Vahdati, S. and Lange, C.: Identifying and Improving Dataset References in Social Sciences Full Texts, *ArXiv e-prints* (2016).

[6] Ikeda, D., Nagamizo, K. and Taniguchi, Y.: Automatic Identification of Dataset Names in Scholarly Articles of Various Disciplines, *International Journal of Institutional Research and Management*, Vol. 4, No. 1, pp. 17–30 (2020).

[7] MacQueen, J.: Some methods for classification and analysis of multivariate observations, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297 (1967).

[8] 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).

[9] 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).

[10] Ohsawa, Y., Liu, C., Suda, Y. and Kido, H.: Innovators marketplace on data jackets for externalizing the value of data via stakeholders' requirement communication, *AAAI Spring Symposium Series-Technical Report*, pp. 45–50 (2014).

[11] Pauline, S. and Jessie, H.: Repositories for research: Southampton's evolving role in the knowledge cycle, *Program: electronic library and information systems*, Vol. 40, No. 3, pp. 224–231 (2006).

[12] Singhal, A. and Srivastava, J.: Data Extract: Mining Context from the Web for Dataset Extraction, *International Journal of Machine Learning and Computing*, Vol. 3, No. 2, pp. 219–223 (2013).

[13] Wong, G. K.: Exploring Research Data Hosting at the HKUST Institutional Repository, *Serials Review*, Vol. 35, No. 3, pp. 125–132 (2009).

[14] 相澤清晴, 松井勇佑, 藤本東, 大坪篤史, 小川徹: 学術漫画データセットの構築〜Manga109〜, 映像情報メディア学会誌, Vol. 72, No. 3, pp. 358–362 (2018).

[15] 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).

[16] 朝岡誠, 林正治, 藤原一毅, 岩井紀子, 船守美穂, 山地一禎: 汎用的データリポジトリにおける制限公開機能の検討と実装, 情報知識学会誌, Vol. 30, No. 2, pp. 168–175 (2020).

[17] 上原直, 早矢仕晃章, 大澤幸生: データ利用者の認識に基づいたデータジャケットの類似度評価, 電子情報通信学会技術研究報告, Vol. 118, No. 453, pp. 23–28 (2019).

[18] 大澤幸生, 早矢仕晃章, 秋元正博, 久代紀之, 中村潤, 寺本正彦: データ市場データを活かすイノベーションゲーム, 近代科学社 (2017).

[19] 大島裕明, 中村聡史, 田中克己: SlothLib: Web サーチ研究のためのプログラミングライブラリ, 日本データベース学会, Vol. 6, pp. 113–116 (2007).

[20] 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).

[21] 小野田崇, 坂井美帆, 山田誠二: 初期値設定法の違いによる k-means 法の性能比較, 第 27 回ファジィシステムシンポジウム講演論文集, No. ME2-4, pp. 231–236 (2011).

[22] 川端隼矢, 長名優子: タッチの類似性を考慮したイラスト検索の精度向上—特徴量と類似度の算出方法の変更—, 情報処理学会第 79 回全国大会講演論文集, No. 1, pp. 31–32 (2017).

[23] 鈴木正敏, 松田耕史, 関根聡, 岡崎直観, 乾健太郎: Wikipedia 記事に対する拡張固有表現ラベルの多重付与, 言語処理学会第 22 回年次大会発表論文集, pp. 797–800 (2016).

[24] 総務省: IoT 時代における ICT 産業の構造分析と ICT による経済成長への多面的貢献の検証に関する調査研究 (2016).

[25] 中川翼, 岡夏樹, 荒木雅弘, 岡海人: 商業施設内における回遊行動履歴を用いた人の属性推定, 人工知能学会第 34 回全国大会論文集, No. 2C5-OS-7b-04, pp. 1–4 (2020).

[26] 日本図書館情報学会用語辞典編集委員会: 図書館情報学用語辞典 第 4 版, 丸善出版 (2013).

[27] 早矢仕晃章, 岩永宇央, 岩佐太路, 大澤幸生: データジャケットを用いた異分野連携に資するデータの特徴とネットワーク分析, 知能と情報, Vol. 31, No. 1, pp. 534–545 (2019).

[28] 久永忠範, 渕田孝康: 統計処理を用いたオープンデータの述語の推薦手法の提案, 情報知識学会誌, Vol. 28, No. 2, pp. 127–133 (2018).

[29] 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).