DEIM2022 E34-4(day3 p26)

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

玄道　俊†　松下　光範††　山西　良典††

† 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: †{k911867,t080164,ryama}@kansai-u.ac.jp

あらまし　本稿では，コンテンツと人の関係に着目した抽象化に基づいて，複数の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．商品検索や料理など様々な目的に応じたアプリケーション研究への利用を念頭に様々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）が公開されており，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）に従ったデータセット単位で整理されている．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ため，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．提案手法では，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する．抽象化によって同一の性質をもつと判断された異なるデータセットに含まれるデータ項目を扱った研究論文について，データ利用に関わる単語の出現傾向を考察することで，提案手法の妥当性を議論した．

キーワード　データセットの整理，データ項目の抽象化，データ整理，学術論文の整理

1　はじめに

商品検索や料理など様々な目的に応じたアプリケーション研究への利用を念頭に，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）が公開されており，それらのデータセットを利用した研究も増加している [7] [6]．例えば，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，2021 年 12 月の時点で 1,143 件あると報告されている [1]．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ことにより，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 [9]．また，新しい手法を考案した場合に，異なる複数のコンテンツデータセットに適用することで，提案手法の汎用性を検証することも可能になる．

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，レシピデータであれば「材料」や「手順」，ホテルレビューデータであれば「宿泊の目的」や「同伴者」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．しかしながら，異なるデータセット間で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．異なるデータセット間で関係性を読み解くためには，複数のコンテンツデータセットに含まれるデータ項目を横断的に把握する必要がある．これは，自らが考案した手法を類似した他の手法と比較する際には特に重要である．

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．共同利用可能なコンテンツデータセットであ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」という二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ことで，それらの活用容易性の向上を図る．

2　データ項目の性質と研究での利用

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．例えば，楽天市場データセットにおける「レビュー内容」と楽天トラベルにおける「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．また，クックパッド

1：国立情報学研究所 HP（https://www.nii.ac.jp）により 2022 年 1 月 11 日確認

データセットにおける「レシピの概要」と楽天市場データセットにおける「商品説明文」は，どちらもコンテンツの特徴を表現する記述であるという共通性が認められる．このように，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., 商品名，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．

意味的に同質の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，そのコンテンツを介してそのコンテンツの利用者や制作者の傾向や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．

3 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．

研究論文中での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 [5]．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．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 [4]．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しきい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．Ayush らや Ikeda ら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．

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 [13]．久永らの研究では，Resource Description Framework(以下，RDF と記す) 形式 [2]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．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 10 社 21 種類 [3]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．

4 提 案 手 法

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，この意味的な共通性として 2 章で述べた「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．

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．例えば，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，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．

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) コンテンツ自身，(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．

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

2：W3C 公式 RDF サイト (https://www.w3.org/RDF/)
3：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) により確認

要素でデータ構造を定義し [13],「＜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．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．

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 1, 図 2, 図 3, 図 4 に記す．図 1，図 2，図 3 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 1，図 3，図 4 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．

5　対象とするコンテンツデータセットの抽象化

5.1　コンテンツデータセットの有向グラフ化

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抽象化する．抽象化するコンテンツデータセットは楽天株式会社 [14] が提供する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，クックパッド株式会社 [11] が提供するクックパッドデータセットの 4 種類とした．

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」をの 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

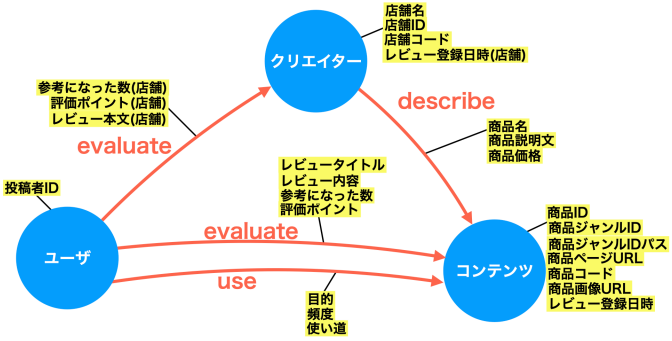

図 1　楽天市場データセットにおけるノードとエッジ図

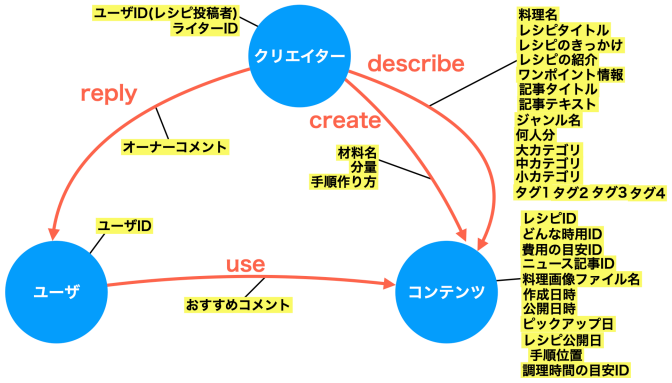

図 2　楽天レシピデータセットにおけるノードとエッジ図

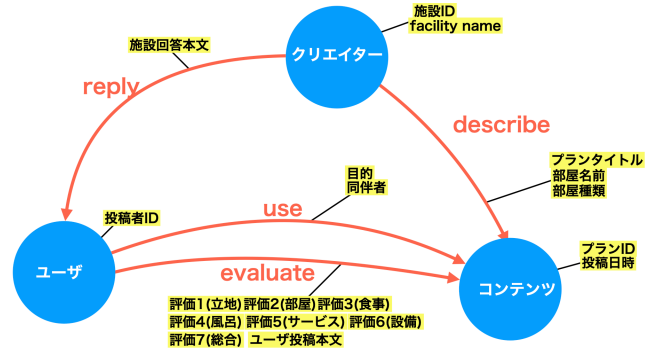

図 3　楽天トラベルデータセットにおけるノードとエッジ図

の有向グラフで表現したものを図 1, 図 2, 図 3, 図 4 に記す．図 1，図 2，図 3 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 1，図 3，図 4 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．

5.2　コンテンツデータセットの共通項目

各コンテンツデータセットの 3 者の関係図から，主語，目的語，述語が共通となるデータ項目（以下，共通項目と記す）を表 1 に記す．（以下，共通項目を表 1 に示すように Ⅰ, Ⅱ, Ⅲ, Ⅳ, Ⅴ, Ⅵ, Ⅶで表す.）

表 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

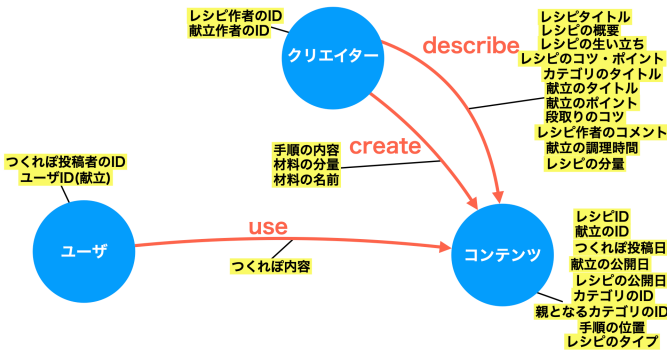

図 4　クックパッドデータセットにおけるノードとエッジ図

6　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．

6.1　検 証 方 法

データ項目の抽象化の妥当性を検証するために，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．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，かつ同一項目を用いた論文が類似しているかを確認するために，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）　論文内における頻出単語の除去．
（3）　出現単語を用い類似性を確認．

各ステップの詳細を以下に示す．

(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は，これらの単語を特定する必要がある．特定するためには，各コンテンツ

データセットを用いた論文でどのような単語が出現したか確認する必要がある．そこで，論文の単語から意味的に分類を行った単語群を作成し，各コンテンツデータセットを用いた論文における単語群の出現回数を算出する．単語群の出現頻度の分散を算出することで，コンテンツごとに単語群の出現の有無を確認できる．例えば，あるコンテンツデータセットを用いた論文には頻出する単語であった場合でも，他のコンテンツデータセットをもちいた論文には頻出しない単語であれば，その単語の頻度の分散は大きくなるつまり頻度の分散が大きい単語群に存在する単語が，コンテンツに依存している単語であると言える．コンテンツに依存した単語を特定するためには，以下のステップを踏む．

(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 件すべての論文の本文情報を形態素解析器 MeCab(ver. 0.996)[4]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[5]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 [8] 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 6 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．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 [12] 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．エルボー法 [3] を用い，クラス数の選定をしたところ 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．各論文の特徴的な単語を選定するために，すべての論文で TF-IDF 値を算出する．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．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 1 を付与した．

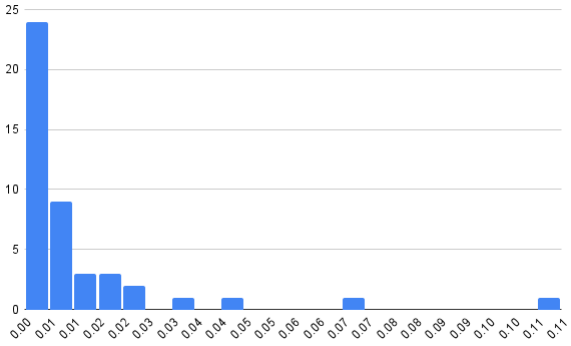

図 5　各単語群における分散

(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，どのクラス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 参照)．横軸は分散値，縦軸はクラス数を表している．図 5 より，分散の値が 0.000 以上から 0.003 未満に集中していることが確認できる．図 5 を概観し，分散の値が 0.004 以上のクラスの分散が高い傾向にあるため，これらのクラス内に含まれる単語を除去することとする．0.004 以上のクラスは 3 つ存在し，「たまねぎ」や「肉」といった料理に関するクラス，「購買」や「小売店」といったショッピングに関するクラス，「ホテル」や「宿泊施設」といった施設に関するクラスであった．料理に関するクラス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 80.0%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) 論文内における頻出単語の除去

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

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

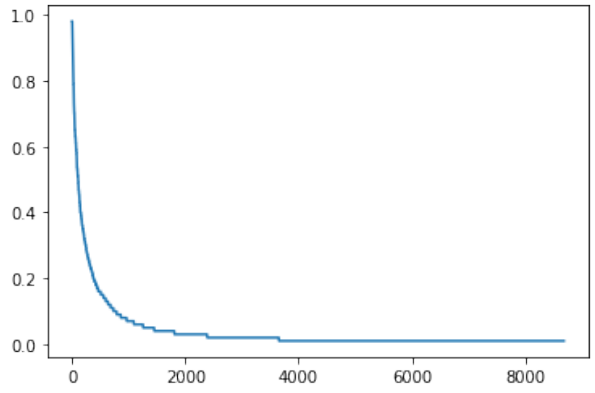

図 6　すべての論文における単語の DF 値

表す DF 値を用い確認する（図 6 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 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] [10]．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 [6]（式 1 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．

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

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．

6.2 検 証 結 果

算出した Bhattacharyya 距離の結果を表 2 に示す．表 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．

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．

7 議　　論

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 [7]．

> 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 [8]．

> 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．

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

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

7：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）

8：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)

表 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	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

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」[9]，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」[10]．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，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．

定性的な調査の結果，共通項目Ⅱ, Ⅳ, Ⅷ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．

8　おわりに

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，Ⅷ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，Ⅶ，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．今後は横断的に把握することができないデータ項目を概観し精度向上を目指す．

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

9：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）

10：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)

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．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」を利用した．記して謝意を表す．

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] 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).

[5] 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).

[6] 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).

[7] 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).

[8] 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).

[9] 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).

[10] 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—, 第 79 回全国大会講演論文集, No. 1, pp. 31–32 (2017).

[11] クックパッド株式会社: 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 (データセット), https://doi.org/10.32130/idr.2.0 (2015).

[12] 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).

[13] 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).

[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), https://doi.org/10.32130/idr.2.0 (2014).