CJ-003

データ形態素：データの意味表現の基本単位探索についての実験的考察
Data Morphemes: Experimental Considerations on the Exploration of Basic Unit of Semantic Representation of Data

早矢仕 晃章†
Teruaki Hayashi

1. はじめに

近年，異なる組織間でデータを交換・流通し，連携することによる価値創出が新たなイノベーションの源泉として注目されている[1, 2]．しかし，異種のデータをやり取りするデータ流通が実現する社会において，データ設計者，取得者，分析者は同一ではなく，異なることがほとんどであり，データの交換と売買は，主に背景知識の異なる他者間で行われる．そのため，データの価値を判断するために提供される情報が不十分であること，そして共通の価値認識を醸成させるための評価指標が存在しないこと等が組織を横断した円滑なデータ取引や流通を困難にしている．データ交換において，事前にデータの価値を正確に見積もることへのニーズはあるものの，どのような情報の提供が有用なのか，またどのような評価指標でデータを評価することがデータ交換を促進するのかということは十分に分かっていない．交換対象のデータの意味を理解するための情報及び評価基準または価値の尺度の確立は喫緊の課題である．

データがビッグであればあるほど意味があるデータであり，利用価値があるのかというと，答えはノーであろう．データの価値付けに対する実社会からの要請はあるものの，価値の試算の難しさには様々な研究で指摘されている[3, 4]．本研究の起点となる問いは，「データの意味や価値は定量的に扱えるのか」ということである．データの意味が数理的に扱えるようになれば，byte といった従来の“量”にフォーカスした方法では捉えられないデータの価値や，背景知識の異なる他者とデータを交換する際の評価基準の策定という“質”の議論を可能となる．

さらにデータの意味や価値を考えていくと，「データがデータとして意味や価値を保持できる限界はどこか」という興味深い問いにたどり着く．言語学における形態素のように，データを構成する要素を細かく分割していくことによって，データたらしめる基本となる構造や単位が存在するのではないだろうか．本研究の究極の目的は，データが意味を持つ基本単位であるデータ形態素の構造とその表現単位の解明である．本論文はデータ形態素の存在を仮定し，データの表現の一つであるメタデータに着目し，データ形態素の探索実験によって得られた結果について考察する．

2. データ形態素

データとは，私たち人間の観察という行為によって取捨選択された実世界の事象の一部を，数字や文字列によって記録したものである（図 1）．まず，私たちは観測したい実世界の事象を選択し，自身の認知モデルに基づいて観察とデータ化を行う．しかし，図 1 に示すように，観察対象のすべての情報がデータになるわけではなく，観察者の目的や能力によってデータ化の段階で様々な情報が喪失する．そして，観察された情報の一部がデータとなる．

実際，この観察とデータ化のプロセスは属人的かつ専門性が極めて高い．そのため，同じ計器で同じ材料を観察しているにも関わらず，筋の良いデータを取る人がいる．例えば，レーザー加工や材料系分野では，観測者のスキルの違いから生じるデータの質が大きく異なる．一方で，スキルが同程度に高い観察者であれば，同じ事象を観測すると同じような品質のデータを取得できる．このようなことから，観察者にある一定水準のデータ取得能力を求めるのであれば，データを意味あるデータたらしめる核となる，数理的に扱える基本構造があると考えることができるだろう．

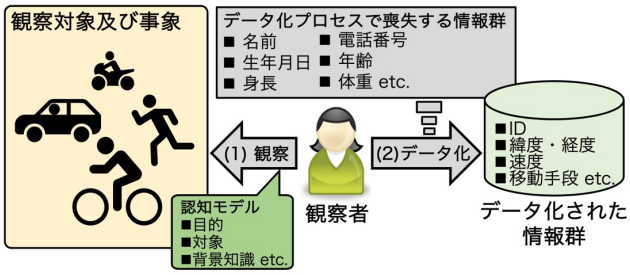

図 1　人間の観察とデータ化プロセスの概念図

データの意味の核となる構造を考える上で，言語学における形態素（morpheme）という考え方を援用してみたい．形態素とは，言葉が意味を有する表現の最小単位を表す．つまり，ある言語においてそれ以上分解したら意味をなさなくなるところまで分割したものが形態素である．例えば，文章は文節，そして単語に分割できる．しかし，単語が必ずしも最小単位ではない．「非科学的」という単語であれば，接頭辞「非」，名詞「科学」，そして接尾語「的」の 3 要素に分けられる．これが意味を有する表現の最小ユニットである．これらの形態素の組み合わせによって，言語のような複雑な意味表現や情報の伝達が可能となっている．

ここで，データにも意味をもつ基本単位が存在すると仮定し，それをデータ形態素（data morpheme）と呼ぶことにする．言語学における形態素の定義に従えば，データ形態素は，それ以上分解したら意味を成さなくなるところまで分割・抽出したデータの表現素子であると言える．

データを意味のある要素に分解していくには，いくつかの段階が考えられる．まず第 1 段階として，データは，実データとメタデータに分けることができるだろう．実データとは，分析の対象であるデータセットそのものであり，数字や文字列によって構成されている．メタデータとはデータのデータを意味する．データの名前や属性，保存形式，保有者などがメタデータである．

†東京大学 大学院工学系研究科 システム創成学専攻
Department of Systems Innovation, Engineering, The University of Tokyo

第 2 段階では，実データとメタデータそれぞれを意味のある構成要素に分解する必要がある．しかし，ここで問題となるのが，表現形態の多様性である．実データは画像データ，表形式データ，時系列データ，音声データなど極めて多様である．また，電子的に扱えるデータもあれば，紙媒体で記録されたデータなど様々である．実データにも形態素が存在するというのが筆者の主張であるが，多種多様な表現形態を持つ実データを統一的に扱うモデルは現在のところ存在しない．一方で，メタデータは DCAT（Data Catalog Vocabulary）[1]や Schema.org[2]などのデータカタログとして標準化されているものが多い．また計算機にとって可読性が高いだけでなく，人間が読んで理解することもフォーカスに入っているため，自然言語で記述されている．そのため実データと比較して，メタデータから意味を抽出することは比較的容易であると考えられる．

そこで本研究では，まずはメタデータを対象にしてデータの意味の基本単位であるデータ形態素の探索を試みる．まず，データ形態素を変数（データに含まれる属性）と仮定してみよう．例えば，「緯度」と「経度」の変数はペアで登場して初めて意味を成す塊となる．「緯度」と「経度」の片方だけでは，位置情報として意味を成さないと言ってよい．そのため，「緯度」と「経度」はペアで初めて位置情報を表すデータ形態素と言えるだろう．また，「住所」という変数はそれ単体で位置情報として機能する．つまり，「緯度」と「経度」のペア，そして「住所」は位置情報のデータ形態素の例ということができる．

しかし，これだけでは何の位置であるのかが不明であり，データの構造は分かっても，当該データの意味を理解するための情報が十分ではない．そこで重要となるのがコンテキストである．例えば，「緯度」と「経度」に「病院」というコンテキストが与えられれば，これらの変数は病院の位置情報であるという意味を成す．つまり，変数はデータ世界において観測対象の事象を記述するために必要な実世界の属性であり，コンテキストは取得する変数群を絞る制約条件である．「緯度」「経度」だけでは大学や会社，人の位置情報かもしれないが，「病院」というコンテキストが制約となり，当該データの意味を定める役割を果たす．

本研究ではメタデータのデータ形態素はコンテキストと変数及びその繋がりによって構成されるという仮定を置く．そして，この仮定に基づき以下の 3 つの構造的特徴を探索し，メタデータにおけるデータ形態素を議論する（図 2）．

① 複数の変数と繋がっているコンテキスト：複数の変数の組み合わせによって意味が定まるコンテキストである．変数が多いということは，観測すべき実世界の属性が多いことを意味している．このようなデータは取得コストが高くなる可能性がある．

② 複数のコンテキストと繋がる変数：様々な文脈に共通して出現する変数は，観測される頻度が高いため，普遍性を有する変数である可能性がある．

③ 特定のコンテキストにおいてのみ登場する変数：このような関係にある変数は，代替可能な他の変数が存在せず，当該コンテキストにおいて重要な変数である可能性がある．

[1] https://www.w3.org/TR/vocab-dcat-3/
[2] https://schema.org/

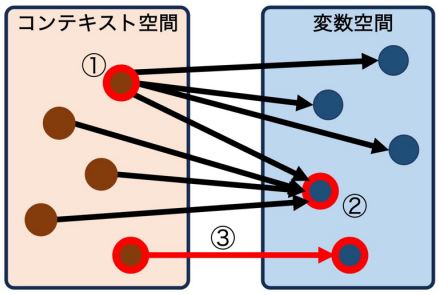

図 2　コンテキストと変数の関係性

3. 実験

3.1 目的

本研究の目的は，データの意味表現の基本単位のモデルを構築するため，実際のメタデータに記述されたデータ表現から意味を構成すると考えられる知識ユニットの構造的特徴を探索することである．本研究では，メタデータの形態素は「コンテキスト」と「変数」によって構成されているという仮定から，実際のデータセットにおけるコンテキストと変数の関係性を分析する．

3.2 データセット

本実験では，データマッチング Web システム TEEDA（Treasuring Every Encounter of Data Affairs）にて収集されたデータアイテム群をデータセットとして用いる．TEEDA は欲しいデータと提供可能データを投稿し，データ提供者と利用者のマッチングを行う Web システムである[3]．データ保有者からの一方的な情報提供のみならず，利用者がどのようなデータを欲しているのかという情報を扱うことでデータ提供者と利用者の邂逅を支援することが可能となる．実験では，TEEDA に公開可能情報として入力されたデータアイテム（データリクエスト：767 件，提供可能データ：295 件）1062 件を分析した（表 1）．

データアイテムの記述形態はメタデータであり，その項目は Schema.org の一部を参照している．本実験ではデータ名（name）と変数（variablesMeasured）を抽出して用いた．コンテキストは簡単のため，データ名に含まれる名詞を対象とした．コンテキストの名詞を抽出する形態素解析には，新語・固有表現に強く，語彙数が多い mecab-ipadic-NEologd[4]を用いた．これにより，「コロナ禍」や「新型コロナウイルス」などの辞書に含まれない単語を的確に抽出した．なお，「データ」と「情報」という名詞はほとんどすべてのデータアイテムに出現する頻出語であったため，ノイズとして除外した．

表 1　データアイテムのサマリー

データアイテム数		1062
コンテキスト	総数	3118
	種類数	1344
	平均保有数	2.94
変数	総数	4717
	種類数	2591
	平均保有数	4.44

4. 結果と考察

4.1 コンテキストと変数の要素数分布

図 3 はコンテキスト及び変数の出現頻度分布を表す．コンテキストでは，「数」「変化」「コロナ」「率」などがある（図 3 (a)）．一方，変数は「年齢」「性別」が最も多く，続いて「人数」「日付」などの出現頻度が高い（図 3 (b)）．「時間」はコンテキストと変数で共通して出現する要素であるが，コンテキストと変数で共通する要素数は 323 件であり，ほとんどの要素はコンテキスト及び変数においてのみ登場する要素となっている．また，コンテキストと比較して，変数の分布はロングテールであり，データの母集団は多数の高出現頻度変数ではなく，多様かつ膨大な低出現頻度変数によって構成されていることが分かる．

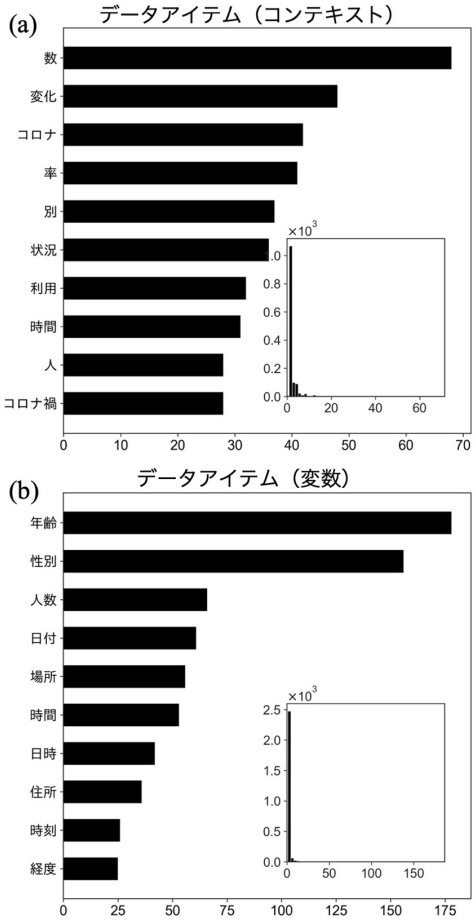

図 3　コンテキスト・変数の出現頻度上位 10 件と出現頻度分布

図 4 はデータアイテムに含まれるコンテキストと変数の数の分布である．ここではデータアイテムにおいて個々のデータが有するコンテキスト数と変数の数に違いがあるのかを検証した．まず，Shapiro-Wilk 検定を適用したところ，コンテキスト数と変数数の対応するデータの差は正規分布であるとは言えなかった．そのため，Wilcoxon 符号順位検定を行った．その結果，個々のデータが有するコンテキスト数と変数の数の平均値に有意な差が見られた（$p < 0.001$）．すなわち，データアイテムが有する変数数はコンテキスト数と比較して多い傾向にあることが分かった．

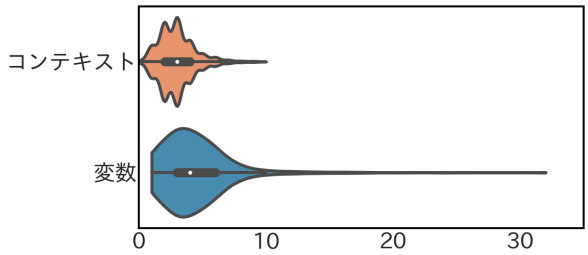

図 4　保有要素数分布
（箱ひげ図の白点は中央値を表し，付帯する曲線は分布を表したカーネル密度グラフを表す）

4.2 コンテキスト-変数のアソシエーションルール

アソシエーション分析とは，データセットから「商品 A を購入した顧客は商品 B も購入している」といったルール（規則性）を抽出する方法である．これをデータアイテム中のコンテキストと変数に適用する．つまり，あるデータにおいてコンテキスト A が現れるとき，変数 B が取得されるといったアソシエーションルール$A \Longrightarrow B$が抽出できる．

なお，アソシエーションルール$A \Longrightarrow B$の支持度（$Sup(A \Longrightarrow B)$）は確率Pを用いて$P(A \cap B)$と表される．アソシエーションルール$A \Longrightarrow B$の支持度が高いということは，全トランザクションにおいてコンテキスト A と変数 B が同時に出現する頻度が高いことを意味する．また，信頼度（$Conf(A \Longrightarrow B)$）は観察者がコンテキスト A において変数 B を取得する条件付き確率であり，$P(B|A)$と表される．

リフト（$Lift(A \Longrightarrow B)$）は以下の式で表され，$Lift(A \Longrightarrow B) > 1$ならば，前提となる条件なしに変数 B を取得したデータと比較して，コンテキスト A において変数 B を取得した割合を表す．つまり，変数 B の取得はコンテキスト A に依存性があることを意味する．

$$Lift(A \Longrightarrow B) = Conf(A \Longrightarrow B)/Sup(B) = P(B|A)/P(B)$$

分析では，まずアプリオリ・アルゴリズムによって支持度 0.005 を閾値として得られた 386 件の集合から，信頼度 0.2 以上かつリフト 3.0 以上で抽出した．その結果，1130 件のルールが得られた．支持度を 0.005 と低い値にしているのは，コンテキスト及び変数が合計 3935 種類と非常に多いためである．続いて，条件部がコンテキストかつ結論部が変数である 85 件のルールを抽出し，図 5 のように可視化した（ノードの大きさはルールにおける出現頻度を表し，エッジの太さは信頼度の大きさを表す）．

「一覧」というコンテキストは，9 種類の変数と繋がっており，他の 16 件のコンテキストと比較して非常に多い．つまり，「一覧」というコンテキストは多くの変数と繋がるコンテキストであり，何かの一覧をデータ化するためには，当該事象の理解に必要な観測変数が多いコンテキストであると言うことができる．また，出現回数が上位のコンテキストでないにも関わらず（図 3 (a)），「一覧」は 85 件のルール出現回数が 52 回と最多であるのも注目すべき点であろう．

一方，異なるコンテキストとよく繋がる変数は「性別」と「年齢」であり，「性別」は 6 つ，「年齢」は 8 つのコンテキストと繋がっている．「性別」と「年齢」は多く出現する変数であるだけでなく，抽出されたルールにおいても多くのコンテキストにおいて観測されやすい普遍的な変数である可能性が示唆される．

最後に注目すべきは，「状況-日時」「企業-企業名」「時間-時間」の 3 つのコンテキスト-変数のペアである．これらは特定のコンテキストとのみよく繋がる変数であり，当該事象の観測に特化した重要な変数の可能性が示唆される．また，これらのペアは信頼度，リフトの値が他のルールと比較して高い．信頼度が大きいということは，コンテキスト A のとき，変数 B が取得される確率が高く，リフトが高いということは，変数 B の取得はコンテキスト A に依存性があると言える．

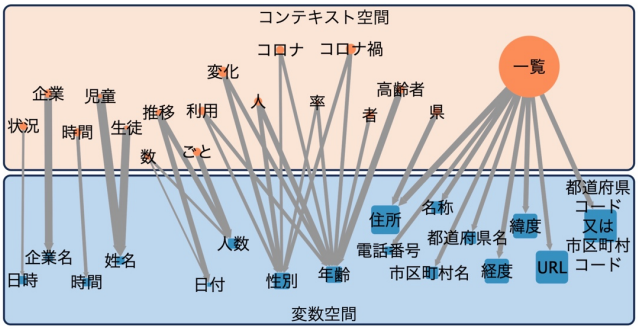

図 5　アソシエーションルールの可視化

4.3 データ形態素についての考察

4.1 節の分析により，コンテキストと比較して変数の出現頻度分布はロングテールであり，データの母集団は多数の高出現頻度変数ではなく，多様かつ膨大な低出現頻度変数によって構成されていることが分かった．また，各データアイテムが持つ変数数はコンテキスト数よりも有意に多い傾向にあり，さらにその種類もコンテキストよりも多いという結果を得た．また，確かに「年齢」や「性別」のような高出現頻度変数が存在するものの，出現回数は高々 178 回であり，変数全体の出現回数のわずか 3.8%である．

また，アソシエーションルールとして抽出されていないものも含めたコンテキスト-変数のペアの出現頻度分布はロングテールな構造となる．12631 ペアのうち，出現回数が 1 回のものが 11638 ペアであり，全体の 92.1%を占める．このことからも，あらゆるデータに出現する普遍的な変数が存在するのではなく，特定のコンテキストに対応して出現する変数が多い可能性が示唆される．

一方で，4.2 節のアソシエーションルールの抽出により，あるコンテキストにおいて取得される可能性の高い変数群の存在が明らかになった．少なくとも，①複数の変数と繋がっているコンテキスト，②複数のコンテキストと繋がる変数，③特定のコンテキストにおいてのみ登場する変数という 3 つのパターンが存在している可能性があり，今後はこれらのパターンを手がかりに，メタデータにおけるデータ形態素の詳細な検討を進める必要がある．

5. 結言

5.1 まとめ

本研究では，言語学における形態素のアナロジーから，データが意味を持つ基本単位であるデータ形態素の構造とその表現の解明を目的とし，データの表現形式の一つであるメタデータに着目して，データ形態素の探索実験を行った．実験では，データマッチングシステム TEEDA で取得されたデータアイテム（提供可能データとデータリクエスト）を分析し，コンテキストと変数の関係からデータの形態素となる要素の探索を行った．

5.2 課題と今後の展望

実験により，データの中で取得される意味の核となるようなコンテキストと変数が存在する可能性が示唆される結果を得たが，本研究ではいくつかの仮定を置いて議論しているため，以下に示す 4 つの大きな課題がある．

まず本研究では，データ形態素の探索においてメタデータを対象としている点である．実データについてもデータ形態素が存在するだろうというのが筆者の主張であるため，今後は多様な実データに対してもデータ形態素となる構造とその表現についても検討する必要がある．

第 2 に，自然言語の扱いである．データ形態素を構成すると考えられる要素であるコンテキストと変数は自然言語で記述されている．コンテキストでは，「COVID-19」と「コロナ」は同じような文脈で利用される単語である．また，「住所」と「所在地」は変数として同じものを表している場合がある．今後の研究では，自然言語処理による要素の類似性を考慮した分析が必要である．

第 3 の課題はデータアイテムの偏りである．TEEDA で収集されたデータアイテムはコロナ禍において注目の集まったデータのメタデータが多く含むため，コンテキストとして「コロナ禍」「コロナ」という語が頻出した．データ形態素についての規則性や普遍的なメカニズムの解明には，より多くのデータアイテムを収集し，分析することが望ましい．また，データアイテムを事前にテーマ等で分類する前処理を行うことも重要であろう．

第 4 に，変数の共起性である．本研究ではコンテキストと変数の関係性のみを対象としたが，変数の中には「緯度」と「経度」のように，同じデータの中で頻出するペアが存在する．このような共起性，変数間の依存関係や階層性を考慮することで変数の代替可能性や変数に特化したデータ形態素の意味表現の特徴を明らかにできるかもしれない．

謝辞

本研究は科研費 JP20H02384 の助成を受けました．

参考文献

[1] F. Schomm, F. Stahl, G. Vossen, “Marketplaces for Data: An Initial Survey,” ACM SIGMOD Record, Vol.42, No.1, pp.15-26 (2013).
[2] 早矢仕晃章, 坂地泰紀, 深見嘉明, “データエコシステムの技術動向とこれからの展開,” 人工知能学会誌, Vol.37, No.5, pp.550-557 (2022).
[3] E.J. Short, S. Todd, “What’s your data worth?” MIT Sloan Management Review, Vol.58, No.3, pp.17-19 (2017).
[4] 高口鉄平, “パーソナルデータは財として扱えるか,” 情報通信政策研究, Vol.4, No.1, pp.47-61 (2020).
[5] 佐藤 敏紀, 橋本 泰一, 奥村 学, “単語分かち書き用辞書生成システム NEologd の運用−文書分類を例にして−,” 自然言語処理研究会研究報告 (2016).
[6] T. Hayashi, Y. Ohsawa, “TEEDA: An Interactive Platform for Matching Data Providers and Users in the Data Marketplace,” Information, Vol. 11, No. 4, 218 (2020).