社団法人　電子情報通信学会
ヒューマンコミュニケーショングループ
第二種時限研究会

語彙の標本化と量子化によるあらすじの特性表現に関する基礎検討

山西 良典[†,a]　　西原 陽子[††,b]　　松下 光範[†,a]

† 関西大学総合情報学部　†† 立命館大学情報理工学部

a) {ryama, m_mat}@kansai-u.ac.jp　b) nisihara@fc.ritsumei.ac.jp

概要　本稿では，人間のコンテンツ認識に関する認知を語彙の標本化と量子化によって模した情報表現を提案し，漫画のあらすじの特性表現への応用を検討する．自然言語処理の分野で様々なタスクに有効性を示している単語分散表現では，連続的なベクトルによって意味が表現されている．一方で，漫画や映画などのコンテンツにおいては離散化された概念の有無によって，コンテンツの特性を捉えていると考えた方が自然なケースが多く見られる．我々は，人間のコンテンツ認識では，連続的な空間での連続的な位置関係を捉えているのではなく，離散的な特性の有無を捉えるアトリビュートのような認識を行っていると考えた．本稿では，この考え方に則ってコンテンツに関わる語彙を標本化し，バイナリに量子化する情報表現手法を提案する．提案手法を漫画のあらすじの特性表現に適用し，その有用性を検討するとともに，漫画あらすじの作品内容の特性とあらすじの書き方の特性を考察した

キーワード　情報表現，コンテンツ特性，漫画検索，物語の特性

1　はじめに

人間はコンテンツの特性を捉えて類似性や相違性を認知するとき，どのようにコンテンツ特性を表現し，どのような演算を行っているのであろうか？例えば，漫画の物語の設定を捉えて他者に作品を紹介するとき，連続的な空間の中に作品を配置し，連続値で表現される距離に応じた作品の特性を解釈することは，一般的な人間には難しい．それよりも，作品の特性を表現可能な離散的概念によって作品の概要を把握し，離散的概念で表現された特性の有無によって表現することが自然ではないかと考える．つまり，例えば漫画のあらすじを表現する上では，

$$\begin{aligned} \text{幽☆遊☆白書} &= \text{能力} + \text{チーム戦} + \underline{\text{妖怪}} \\ \text{烈火の炎} &= \text{能力} + \text{チーム戦} + \underline{\text{忍者}} \\ ONEPIECE &= \text{能力} + \text{チーム戦} + \underline{\text{海賊}} \\ \text{バジリスク} &= \text{能力} + \text{チーム戦} + \underline{\text{忍者}} + \underline{\text{悲恋}} \end{aligned}$$

のように，離散的な概念の組み合わせ[1]によって，物語の設定を捉えることが一般的であると考える．

自然言語処理の研究分野では，word2vec [1] や fastText [2]，BERT [3] といった連続値で表現された多次元ベクトルによる情報表現が，検索，翻訳，分類などの様々なタスクにおいて高い性能を示している．人間がある概念を捉えるとき，一つ一つの単語の意味などは連続的な空間の中での距離を把握していると考えることは直感的にも妥当であり，一般的な言語処理に関するタスクで高い性能を示すことには納得がいく．しかしながら，人間の直感や感性によって享受されるコンテンツを対象とした場合には，必ずしも連続的な表現が上手く働くとは限らず，連続値の空間を標本化したうえで捉えていると考える．上記の例で言えば，「幽☆遊☆白書」と「烈火の炎」，「ONE PIECE」では，それぞれ「妖怪」，「忍者」，「海賊」という特性が異なるのみであるが， 連続値空間の中での位置として表現した場合には作品自体を表現するベクトルの位置は，「妖怪」「忍者」「海賊」それぞれの単語の全体空間中での距離に応じて異なることになる．「バジリスク」は，「烈火の炎」にロミオとジュリエットのような「悲恋」の要素が加わったとみなせるが，ベクトルの平均化処理によって，悲恋という要素が他の特性に薄められてしまっては，物語の本質を表現できるとは言えない．

人間がコンテンツの特性を捉える上での概念化についての問題もある．例えば，あらすじに「鬼」「幽霊」「妖怪」「ドラキュラ」といった単語が複数出現したとする．この場合，人間の直感ではこれらの単語がそれぞれ何回出現するかではなく，単純に「鬼や妖怪などの異形に関すること」が出現する物語であるか否かをバイナリで評価する方が自然な情報表現であるのではないかと考える．特定の表現やイベントを忌避する場合 [4] には，頻度が少なかったとしても，特性が発生する時点でコンテンツに対する楽しみが減少してしまう可能性がある．

本稿では，人間が離散的な概念の有無によってコンテンツ特性を把握していると仮定する．この仮定にもとづいた特性表現手法として概念アトリビュートを提案し，漫画のあらすじの特性を情報表現する．概念アトリ

[1]ただし，これらは例示のための第一著者の解釈である．

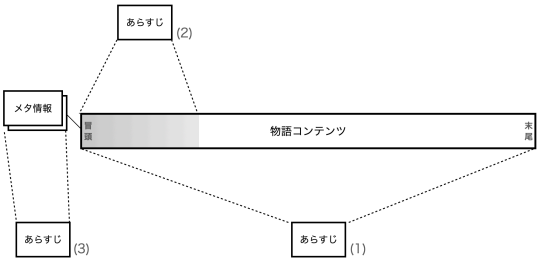


図 1　本文とあらすじの関係．(1) 全体を端的に要約したあらすじ，(2) 作品の世界観が伺える記述，(3) 作品の社会的受容に関する言及．

ビュートによる漫画の物語の特性表現への有用性の検証と，漫画のあらすじに特質的に現れる記述の特性について基礎的な考察を実施する．

2　漫画のあらすじ

一般に漫画のあらすじと言えば，コンテンツ全体の要約を想像するかもしれない．例えば,「ONE PIECE」1 巻のあらすじとして期待するのは，1 巻全体を端的に要約したもの（(図 1-(1)）であろう．しかし，実際に電子漫画配信サイト等から取得できるあらすじは，ネタバレ防止の観点から当該巻の全体の要約にはなっておらず，その冒頭部の内容や登場人物のプロファイルといった，いわゆるその作品の「世界観」が伺える記述（図 1-(2)）であったり，それに「アニメ化された」や「大人気クリニックコメディ」のような，その作品の社会的受容に関する言及（メタコンテンツ）を付与したもの（図 1-(3)）であったりする．また，(2) と (3) が複合されたあらすじもしばしば観察される．これは電子漫画配信サイトに掲載されるあらすじは，作品の概要を伝えることのみならず，漫画の販売促進や購入検討者の誘引を企図したものであることに依る．

電子漫画配信サイトでは作品にタグが付与される場合が多く，それを用いればあらすじに用いられる単語を概念化する必要がない，という指摘が想定される．そこで，電子漫画配信サイトで実際に付与されているタグについて調査した．一例として eBookJapan にて「医療」というタグが付いている作品を取り上げる．当該サイトにてこのタグが付与された 1,943 件のあらすじをスクレイピングし，これらのテキストを精査した結果，

1. 医療漫画ではないにもかかわらず，単に主人公が「医者」だけで,「医療」タグが振られているものがある（e.g., きるる KILL ME）
2. 獣医漫画にも「医療」タグが振られている
3. 医療の専門的な事象がメインの漫画と，医療にまつわる人間ドラマがメインの漫画がある．

といった事例が確認された．このうち，1 は，付与されているタグが必ずしもジャンルを表しているものでないことを端的に示している．また，2 はタグが示す概念の粒度が，読者が想定するものとは必ずしも一致しない可能性を示唆している．これらを適切に識別可能な概念レベルでの情報表現が可能になれば，読者の選書体験が向上すると期待される．

これまでに，物語構造を分析・応用した研究は存在する（例えば，文献 [5]）．一方で，あらすじの構造の分析，特に電子漫画配信サイトのような商業的側面が多いドメインにおけるあらすじの特性については，著者らの知る限り少ない．

3　概念アトリビュート

漫画のあらすじを分析する上で，人間の認知に適応してコンテンツの特性を認知するための情報表現方法として概念アトリビュートを提案する．本章では，概念アトリビュートの提案に至った背景となる考え方と概念アトリビュートの獲得方法を述べる．

3.1　背景となる考え方

計算機によるコンテンツの認知では，人間の認知に合致させるための前処理として，連続的な物理特徴量の標本化が行われてきた．例えば，聴覚の認知においては，周波数分析によって得られた各周波数の振幅に対して，人間の聴覚特性 [6] に基づいて設定された窓関数（メルフィルタバンク）を用いて周波数帯域を分割し，それぞれの周波数帯域ごとの振幅を特徴量として用いることが一般的である．物理特徴量として得られた周波数特徴量は，人間の音に対する認知を反映して標本化された特徴量に変換した上で，音声の発話認識や感情認識が行われている．音楽情報処理では更に上位の概念で標本化し，西洋音楽で用いられる 12 音階に周波数帯域を離散化したクロマベクトル [7] も提案されている．440Hz と 880Hz は倍音の関係にあり，これは音楽的にはどちらも A の音として認識される．例えば，{A，C♯，E} という A major の和音認知においては，A の音としては 440Hz と 880Hz のどちらが再生されていても,「A の音が存在する」というバイナリに量子化された認知に従ってどちらも A major の和音として認知される．クロマベクトルは，このような音楽的な意味に従って複数の周波数帯域をまとめることで，和音認識などのアプリケーションに利用されている．人間の画像に対する認知を模した物体認識でも，意味的に標本化された属性の有無を示すアトリビュートによって，説明可能な特徴量を用いながらもゼロショット物体認識 [8] で有効な性能を示している．これらの先行事例は，コンテンツ認識における物理特徴量に対する標本化の重要性を示唆していると考える．

自然言語処理研究において一般的に用いられる単語分散表現は，単語というシンボルを単語の文脈的な特性に基づいて連続値ベクトルで特徴表現する技術であり，それ自身では上記で述べた連続値の特徴量空間の標本化は行われていない．したがって，単語分散表現そのものから統計特徴量を算出したとしても人間の意味的なコンテンツ認識とはかけ離れた表現となっている可能性が考えられる．語彙の概念化に関する研究としては，英略称の推定 [9] や多義語の文脈ごとの分散表現の獲得 [10] などが報告されている．しかしながら，これらはある単語に対する複数の概念を対象としたものであり，複数の単語によって構成される上位の概念によるコンテンツの理解を目指したものではない．コンテンツ認識には，複数の単語の集合から構成される「概念」を表現する方法が必要と考えた．

3.2 概念アトリビュートの獲得

3.1 節の考え方に則って，コンテンツの特性を離散的，量子的に捉える情報表現として，概念アトリビュートを提案する．概念アトリビュートは，特定のドメインにおける語彙を獲得し，意味的に近い単語をクラスタとしてまとめることで得られる．特定のドメインで扱われる語彙全体をある一定の粒度で意味的に近い単語集合に標本化し，この概念アトリビュートの有無によってコンテンツ特性の認識をねらう．

概念アトリビュートは，以下の手順で得られる．

1. ドメイン内の一般語彙の獲得
2. 語彙中の単語についての単語分散表現ベクトルの獲得
3. 単語分散表現ベクトルを用いた分類による語彙の標本化
4. 対話的処理による概念アトリビュートの精錬

これらの手順のうち，手順 3 までについては，Bag-of-features に基づく画像処理技術のひとつである SIFT 特徴量 [11] での Visual-word の獲得までの流れとほぼ等しい．SIFT 特徴量において，局所特徴量に当たる部分が各単語の単語ベクトルとなり，局所特徴量の分類で得られる Visual-word が語彙の標本に相当する．Visual-word とは異なり，人間が意味を理解可能な単語の集合として語彙の標本が得られるため，手順 4 で対象ドメインの特性表現においてノイズとなる標本を取り除く必要がある．本稿では，手順 3 までの処理を漫画のあらすじに適用し，漫画のあらすじの特性分析における概念アトリビュートの潜在的な可能性を議論する．

3.2.1 ドメイン内の一般語彙の獲得

大規模データを参照して，分析対象ドメインで扱われる単語を語彙として獲得する．例えば，様々な漫画作品のあらすじを対象とした場合には，掲載雑誌や漫画ジャンルなどにも網羅性が高い大量のあらすじのデータを用意する．あるいは，医療漫画の中での違いを分析したいと言った場合には，医療漫画のみのあらすじを収集したデータセットを用意する．

概念アトリビュートでは，複数のコンテンツの特性の共通点や相違点を一覧可能にすることを狙いとしている．そこで，一定数のコンテンツで共通して用いられる単語のみを概念アトリビュートを構成する語彙とする．得られた全ての単語に対して文書集合中での Document Frequency（df）を算出する．分析対象とした文書数中の df 値の割合を算出し，これを相対 df 値とする．任意の単語の相対 DF 値でロングテールに該当する単語については語彙から除外することで一般語彙を獲得する．これは，「複数のコンテンツにおいて共用可能である」というアトリビュートの根本的な性質の条件に従うものである．

本稿では，ストップワードを除く名詞のみを対象として一般語彙を獲得することとした．用意した大規模データの中から，ドメイン中で扱われている単語を獲得する際には，品詞に絞って語彙を獲得することによって有用な概念アトリビュートが得られる場合もある．例えば，名詞のみを扱うことで直接的に概念を端的に捉えた方が良い可能性もある一方で，ドメインによっては形容詞や形容動詞といった性質を表現する品詞も対象として語彙を獲得した方がコンテンツの特性をより適切に表現可能である可能性もある．コンテンツの特性を端的・離散的に表現する上で，どのような品詞が適切な素性となり得るかについては議論が必要であるが今後の課題とする．

3.2.2 単語ベクトル分類による概念アトリビュートの獲得

3.2.1 節で獲得したドメイン内の一般語彙の各単語について，単語分散表現ベクトルをそれぞれ得る．このとき，各単語の単語分散表現については，大規模語彙を対象としてあらかじめ学習された事前学習済みモデルを参照して獲得すれば，より一般的な単語に対する認識において語彙中の単語を評価することになる．一方で，対象ドメインの語彙を獲得した大規模データから学習して得られた単語分散表現モデルを用いることも可能である．どのような視座からコンテンツを捉えて素性を表現するのかによって，単語分散表現モデルは異なるものを参照すべきであると考えられる．本稿では，一般的な大規模語彙を事前に学習した事前学習モデルを用いる．参照する単語分散表現モデルについては，コンテンツの特性や

アプリケーションとしての有用性などに応じて議論が必要である．

単語分散表現ベクトルによって表現された各単語に対して，任意のクラスタ数 k を指定し，k-means++ [12] で分類する．これにより，語彙中の単語について意味的な類似性が高い単語集合が一定の概念に分類・整理されることを期待する．つまり，類似した意味を持つ単語同士は同じクラスタとしてまとめて扱うことで，人間の認知に近い感性でドメインの語彙を解釈可能になることをねらう．このとき，任意のクラスタ数 k については，様々な決定方針が考えられる．例えば，クラスタ内誤差平方和が最小となるクラスタ数を探索的に求める方法や，*X-means* [13] や *VBGMM* [14] のようにクラスタ数を自動的に決定する方法を用いることも選択肢として挙げられる．本プロセスにおける k の決定方法については今後の議論とする．

3.2.3 対話的処理による概念アトリビュートの精錬

3.2.2 節で得られた概念クラスタそれぞれに対して，対話的処理によって精錬化を行う．精錬化では，ヒューマンコンピューテーションによって，各概念クラスタがドメインの特性を示す単語の集合として適切であるかについて評価する．

どのようなインタラクションデザインで一般的にコンテンツの特性を表現可能な概念アトリビュートを獲得するのかについては，精錬前の状態でどのような単語が群化されるかを分析したうえでの検討が必要となる．そこでまず，本稿では，概念アトリビュートの精錬前の状態で，漫画のあらすじの特性を表現可能な単語集合がまとまって得られているのかについて検討する．

4 漫画のあらすじ特性表現のための概念アトリビュートの獲得

漫画のあらすじを対象として，提案手法にしたがって概念アトリビュートを獲得し，あらすじの特性を分析する．なお，本稿では生成結果に対する統計的データと生成された概念アトリビュートへの考察のみとし，複数の評価者による客観的な精錬と評価については今後の課題とする．

4.1 対象データ

eBookJapan [2]において,「異世界タグ」が付与された漫画 2,099 作品（以下，ファンタジー漫画）と「医療タグ」が付与された漫画 1,943 作品（以下，医療漫画）を抽出した．これらの各作品のあらすじを抽出し，分析に用いた．ストップワードの設定には，SlothLib[15] を利用した．

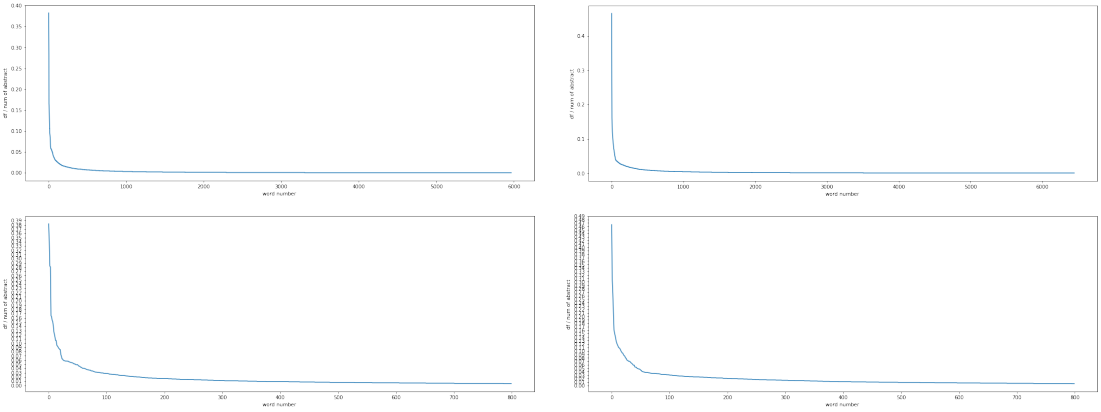
ファンタジー漫画と医療漫画の語彙に含まれる単語の種類数は，それぞれ 5,966 単語と 6,451 単語となった．ファンタジー漫画は医療漫画に比べて若干作品数が多いにも関わらず，あらすじで用いられる単語の異なり数としては医療漫画の方が多い結果となった．表 1 に，各ジャンルのあらすじでの相対 df 値上位 20 件の単語とその相対 df 値を示す．形態素解析の問題で，ストップワードとしてフィルタリングされなかった記号列がいくつか見られる．本来は，これらの単語は概念アトリビュートを構成する単語ではないため除外すべきではあるが，本稿ではこれらを含めた考察を行う．

表 1　ファンタジー漫画と医療漫画それぞれのあらすじ集合での DF 値上位 20 件の単語．表中では，単語とその相対 DF 値を示す．相対 DF 値は小数点 4 位以下を切り捨てとした．あらすじとしてのメタ的な特性を示すと考えられる単語については下線を付した．

順位	ファンタジー漫画	医療漫画
1	('!!', 0.382)	('!!', 0.465)
2	(' 世界', 0.340)	(' 患者', 0.303)
3	(' 転生', 0.282)	('収録', 0.265)
4	('収録', 0.280)	(' 病院', 0.209)
5	(' 魔法', 0.165)	(' 手術', 0.160)
6	(' 最強', 0.162)	(' 医師', 0.151)
7	(' 勇者', 0.154)	(' 命', 0.144)
8	('ファンタジー', 0.151)	(' 医療', 0.135)
9	(' 召喚', 0.141)	('?', 0.126)
10	(' 魔王', 0.126)	(' ナース', 0.121)
11	('...', 0.120)	(' 事件', 0.117)
12	('?', 0.113)	(' 臓器', 0.113)
13	(' 人間', 0.106)	(' 心', 0.109)
14	(' 少女', 0.106)	('...', 0.108)
15	(' 魔物', 0.095)	('作品', 0.1070)
16	(' スキル', 0.093)	('...!?', 0.103)
17	(' ドラゴン', 0.091)	(' 治療', 0.099)
18	(' 冒険', 0.088)	(' 謎', 0.097)
19	(' 能力', 0.087)	(' 売買', 0.094)
20	(' 人生', 0.0852)	(' 少女', 0.093)

どちらのジャンルにおいても相対 DF 値の第 1 位は “!!” であった．これは，作品の内容自体を表現する単語ではないものの，あらすじにおいて多用される記号列として抽出されたものと考えられる．この他にも，“?” や “...” などの記号列もどちらのジャンルでも相対 df 値の上位として抽出されており，ジャンルによらずあらすじの記述に利用される記号列であることがわかる．他にも，“収録” はどのような話が単行本に含まれているかを示すメタ的な単語であり，内容そのものを表すわけではない．メタ的な単語を除くと，ファンタジー漫画では，“世界” “転生” といった単語が全体の 30%前後で用いられている．医療漫画では，“患者” “病院” という単語は 20%を超えているが，その他の単語はおおよそ 15%以下の相対 df 値となっている．2 種類のジャンルを比べてみる

[2] https://ebookjapan.yahoo.co.jp/



(a) ファンタジー漫画　　(b) 医療漫画

図 2　ファンタジー漫画と医療漫画それぞれにおける単語の相対 df の分布．上図は語彙全体を下図は上位単語部の拡大図を示す．

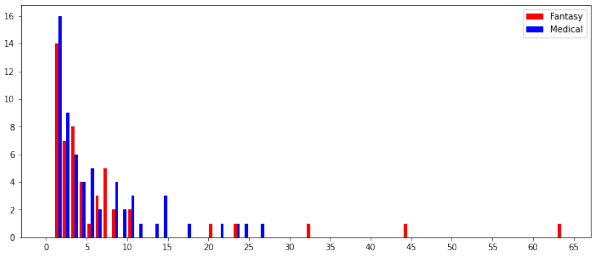


図 3　ファンタジー漫画と医療漫画それぞれのクラスに含まれる単語数のヒストグラム

と，ファンタジー漫画には医療漫画には出現しない “魔法” “召喚” といった単語が，医療漫画にはファンタジー漫画には出現しない “臓器” “手術” といった単語がそれぞれ出現しており，ファンタジー漫画と医療漫画で作品で描かれる内容の違いが表現されていることがわかる．

図 2 に，ファンタジー漫画と医療漫画それぞれにおける単語の相対 df の分布を示す．同図から，どちらのジャンルの漫画のあらすじにおいても，複数の作品で共通して用いられる単語は，あらすじで用いられる総単語から考えると限定的であることがわかる．ファンタジー漫画と医療漫画のどちらでも，相対 df 値が 0.01 を下回った辺りからロングテールとなっている．

4.2　概念アトリビュートの獲得結果

図 2 で示した語彙の相対 df 値の分布から，一般語彙抽出のための相対 df のしきい値を 0.01 に設定した．その結果，ファンタジー漫画では 360 語，医療漫画では 440 語がそれぞれ抽出された．これらの単語それぞれに対して，大規模語彙で事前学習された単語分散表現ベクトルを獲得した．本稿では，Wikipedia をベースとして学習されたモデル[3] [16] を用いて単語分散表現ベクトルを得ることとした．その結果，ファンタジー漫画では 345 語，医療漫画では 427 語の単語ベクトルが得られた．単語ベクトルが獲得できなかった単語は，記号列やキャラクタ名などの固有名詞であった．単語分散表現ベクトルが獲得できた単語群をそれぞれのドメインでの概念アトリビュートを構成する素子として扱う．

概念アトリビュートのクラスタ数 k については，人間の複数オブジェクトへの認知限界として知られている Magical number [17] を参考に，各クラスタに含まれる平均単語数が 7 となるように設定した．結果として，ファンタジー漫画では 51 個，医療漫画では 62 個の概念アトリビュートが得られた．図 3 に，それぞれのドメインでクラスタあたりに含まれる単語数のヒストグラムを示す．どちらのドメインでも 1 単語のみで構成されるクラスタが最も多く生成されていることがわかる．ファンタジー漫画では 30 以上の単語で構成されたクラスタがいくつか存在している．それ以外は，5 単語前後で構成されていた．2 単語以上で構成されるクラスタ数はファンタジー漫画と医療漫画のそれぞれで，36 クラスタ（全 51 クラスタ中）と 46 クラスタ（全 62 クラスタ中）であった．これらのクラスタに着目して，あらすじや漫画の内容を理解する手がかりとなる概念としてまとめられているのかについて次節で考察する．

4.3　考察

本来の概念アトリビュートの獲得としては，最終的に対話的処理によってノイズとなるクラスタなどを取り除いて精錬化していく．本稿では，著者ら 2 名が行った精錬結果を示し，概念アトリビュートによるあらすじの特性表現について考察する．

全体としては，ファンタジー漫画，医療漫画のどちらでも作品のメタ的な特性を示す単語群と物語の内容を示す単語群に分類され，物語の内容を示す単語群では単語

[3] http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/

表 2　ファンタジー漫画のあらすじに見られたメタ特性を示す単語集合．説明の便宜上 ID を示す．

ID	単語集合
M_F1	’イラスト’, ’小説’, ’原作’, ’作品’, ’コミック’, ’作者’
M_F2	’内容’, ’サイト’, ’差異’, ’巻末’, ’デジタル’, ’単行’, ’扉’, ’過去’, ’クラス’, ’画像’, ’全て’, ’ランク’, ’おまけ’, ’電子’, ’累計’, ’シリーズ’, ’トラック’, ’完結’, ’web’, ’タイトル’
M_F3	’商品’, ’価格’, ’サービス’
M_F4	’収録’, ’連載’, ’掲載’
M_F5	’ページ’, ’巻’

表 3　医療漫画のあらすじに見られたメタ特性を示す単語集合．説明の便宜上 ID を示す．

ID	単語集合
M_M1	’開始’, ’開幕’, ’スタート’
M_M2	’話題’, ’人気’
M_M3	’作品’, ’シリーズ’, ’名作’, ’傑作’, ’作家’
M_M4	’漫画’, ’コミック’, ’著者’, ’雑誌’, ’マンガ’, ’作者’, ’編集’, ’タイトル’, ’ビデオ’
M_M5	’物語’, ’舞台’, ’ミステリー’, ’完結’, ’原案’, ’ドラマ’, ’サスペンス’, ’前作’, ’ギャグ’, ’ヒーロー’, ’エピソード’, ’描写’, ’ホラー’
M_M6	’収録’, ’掲載’, ’連載’

が示す物語の特性に分割されていることが確認できた．また，あらすじの特性としてはノイズと見られる単語についても，ノイズクラスタとして集約されていることを確認した．一方で，1 クラスタ内に大量の単語が無秩序に分類され，あらすじの特性としては判別不可なクラスタも存在した．

4.3.1　メタ特性

分析対象とした 2 単語以上から構成されるクラスタについて，ファンタジー漫画と医療漫画ではそれぞれ 5 クラスタと 6 クラスタがメタ特性を示すと評価できた．表 2 に，ファンタジー漫画のあらすじのメタ特性として評価されたクラスタの単語集合を示す．M_F1 はトランスメディアに関連する単語群，M_F2 と M_F3 は広告用の単語群，M_F4 と M_F5 は書誌情報に関連する単語群がクラスタとして得られていることがわかる．表 3 に，医療漫画のあらすじのメタ特性として評価されたクラスタの単語集合を示す．M_M1 は物語の始まりを示す単語群，M_M2 と M_M3 は作品の評判，M_M4 や M_M5 は作品やメディアをメタ的に示す単語がそれぞれクラスタとして得られていることがわかる．また，M_M6 は M_F4 と全く同じ単語集合で構成されており，ジャンルを問わずにあらすじで用いられるメタ的な概念アトリビュートであると推察される．

これら 2 種類のジャンルであらすじに記述されるメタ特性の違いに着目したい．ファンタジー漫画では，電子マンガ特有と見られる広告用の単語が集まったクラスタが得られた．また，M_F1 のように他メディアとの関連を示すクラスタも得られた．一方で，医療漫画では M_M3 や M_M4 のように作品をメタ的に示す単語は得られたものの，電子配信に関連する単語群や他メディアに関連する単語群は見られない．これは，ターゲット読者層の違いや電子配信されるまでの流れの違い（始めから電子的に配信 or 紙面で出版後に電子配信）によるものである可能性があると考えられる．これらは，漫画の内容を直接的に指し示すものではないが，作品の周辺情報（読者やメディアなど）を示す重要なあらすじ特性であると考えられ，概念アトリビュートとして利用できると考える．

表 4　ファンタジー漫画のあらすじに見られた内容特性を示す概念アトリビュート．説明の便宜上 ID を示す．

ID	単語集合
C_F1	’魔法’, ’モンスター’, ’敵’, ’ステータス’, ’種族’, ’バトル’, ’呪文’, ’武器’, ’キャラ’, ’アイテム’
C_F2	’精霊’, ’姫’, ’女神’, ’神’, ’女王’, ’王女’, ’魔女’
C_F3	’美少女’, ’主人公’, ’女の子’, ’美女’, ’女性’, ’ヒロイン’, ’悪役’, ’人物’
C_F4	’貴族’, ’英雄’, ’騎士’, ’奴隷’, ’紋章’, ’戦士’, ’族長’
C_F5	’毒’, ’料理’, ’肉’, ’腹’
C_F6	’事故’, ’事件’, ’危機’, ’戦争’, ’事態’, ’革命’, ’行動’, ’襲撃’, ’トラブル’, ’騒乱’

表 5　医療漫画のあらすじに見られた内容特性を示す概念アトリビュート．説明の便宜上 ID を示す．

ID	単語集合
C_M1	’先生’, ’メス’, ’動物’, ’猫’, ’卵’, ’犬’, ’部屋’, ’ママ’, ’飼い主’
C_M2	’教授’, ’博士’, ’助手’, ’会長’, ’部長’
C_M3	’死体’, ’解剖’, ’調査’, ’遺体’, ’検査’, ’チェック’, ’研究’, ’死因’
C_M4	’売買’, ’現場’, ’目的’, ’組織’, ’解体’, ’標的’, ’狩り’, ’ビジネス’, ’日常’, ’刑務所’, ’生活’, ’体験’, ’直接’, ’対策’, ’放火’, ’住人’, ’売春’
C_M5	’心’, ’日々’, ’人生’, ’笑顔’, ’想い’, ’悩み’, ’夢’, ’恋’, ’友情’, ’気持ち’, ’一生’, ’思い’, ’愛’, ’絆’
C_M6	’家族’, ’母親’, ’出産’, ’母’, ’妊娠’, ’父親’, ’親’, ’両親’
C_M7	’臓器’, ’原因’, ’リスク’, ’症状’, ’妊婦’, ’移植’, ’ウイルス’, ’薬’, ’ストレス’, ’感染’, ’腎臓’

4.3.2　内容特性

ファンタジー漫画では 28 クラスタ，医療マンガでは 36 クラスタがそれぞれ内容を示す概念アトリビュートとして評価された．表 4 と表 5 に，いくつかの代表的なクラスタを例示して考察する．

まず．表 4 に示したファンタジー漫画の内容特性を示す概念アトリビュートを考察する．C_F1 はロールプレイングゲームに出現するような要素を示す単語集合が得られている．隣接領域であるゲームのような展開が描かれる漫画であるかどうかを判別するうえで，この概念アトリビュートは有効に働くと考えられる．C_F2 と C_F3 は女性のキャラクタ属性，C_F3 はキャラクタのプロファイル属性を示すような単語，C_F4 は料理に食事に関連する単語集合，C_F5 は物語中に発生するあまりポジティブではないイベントを示す単語がおおよそ集合として，それぞれ得られている．C_F2 と C_F3 は結合して女性キャラクタに関するクラスタとして得られてもよいが，C_F3 が一般的な女性を示す単語である一方 C_F2

表 6　ファンタジー漫画のあらすじに見られた概念を判定不可なクラスタ．説明の便宜上 ID を示す．

ID	単語集合
Q_F1	’ファンタジー’, ’街’, ’ライフ’, ’プロジェクト’, ’ギルド’, ’住人’, ’サラリーマン’, ’家族’, ’職業’, ’グルメ’, ’学園’, ’令嬢’, ’パーティー’, ’仕事’, ’人達’, ’学校’, ’中年’, ’異色’, ’コメディ’, ’日常’, ’パーティ’, ’夜’, ’都市’, ’カフェ’, ’メイド’, ’マジック’, ’カー’, ’会社’, ’コメディー’, ’モノ’, ’舞台’, ’世界中’
Q_F2	’開幕’, ’転移’, ’待望’, ’無双’, ’!!!’, ’現代’, ’今日’, ’ヒナ’, ’ハーレム’, ’エッチ’, ’幼女’, ’奮闘’, ’社畜’, ’...?’, ’リムル’, ’過労’, ’寝間’, ’招喚’, ’存続’, ’協力’, ’オーク’, ’依頼’, ’女子’, ’最初’, ’満足’, ’モテ’, ’討手’, ’招待’, ’満喫’, ’念願’, ’白虎’, ’触り’, ’!「’, ’仕込み’, ’場所’, ’クソ’, ’川’, ’ステーキ’, ’ーー’, ’!?(’, ’こころ’, ’童貞’, ’種付け’, ’最中’, ’増量’, ’分冊’, ’通り魔’, ’里’, ’アキト’, ’最終’, ’導’, ’おっさん’, ’途中’, ’遺跡’, ’マイペース’, ’サバイバル’, ’旅路’, ’異種’, ’ハルト’, ’!!?’, ’試し’, ’今度’, ’希代’, ’クロ’

表 7　医療漫画のあらすじに見られた概念を判定不可なクラスタ．説明の便宜上 ID を示す．

ID	単語集合
Q_M1	’?’, ’ナース’, ’,「’, ’倉’, ’環’, ’裏’, ’天’, ’朝顔’, ’女医’, ’くるみ’, ’待望’, ’琢磨’, ’がん’, ’ひがし’, ’更正’, ’ブライダル’, ’渾身’, ’筆致’, ’禁断’, ’?」’, ’鬼才’, ’)=’, ’折’, ’知’, ’フロンティア’, ’たま’, ’金字塔’, ’¿’, ’:「’, ’整形’, ’ナス’, ’こころ’, ’天使’, ’所収’, ’検案’, ’功’, ’蓮’, ’...?’, ’深奥’, ’なき声’, ’仁’, ’倫’, ’最中’, ’(「’, ’けい’, ’うつ’, ’アダルトビデオ’, ’...「’, ’)」「’, ’雑学’, ’途中’, ’!】’, ’ガン’, ’モノ’, ’エッチ’, ’クジラ’, ’¡’, ’義手’, ’影’, ’響’, ’不妊’, ’蝶’, ’リュウ’, ’天道’, ’難事件’
Q_M2	’本’, ’過去’, ’カラー’, ’全て’, ’期待’, ’累計’, ’電子’, ’数々’, ’時代’, ’歴史’, ’最後’, ’最終’, ’現代’, ’今日’, ’最新’, ’最先端’, ’同時’, ’現在’, ’グループ’, ’当時’, ’最前線’, ’幕末’, ’紙’, ’前後’, ’チーム’, ’後半’
Q_M3	’闇’, ’人間’, ’自ら’, ’価値’, ’欲望’, ’社会’, ’世界’, ’未来’, ’肉体’, ’現実’, ’理想’, ’行動’, ’絶対’, ’個性’, ’モンスター’, ’同士’, ’闘い’, ’生命’, ’想像’, ’心理’, ’正義’, ’究極’, ’魔法’

は役職やジョブのような単語が集まっており，同じ女性キャラクタに関する単語であっても紐付く概念の違いに応じたクラスタが形成されている．これらの概念アトリビュートを使い分ければ，例えば「王女のような強い立場の女性出てくる物語」を読みたいと考えたとき，一般的な女性キャラクタが出現する物語と差別化して捉えることができると考えられる．

つぎに，表 5 に示した医療漫画の内容特性を示すアトリビュートについて考察する．C_M1 は，動物に関わる単語が集まっており，医療漫画の中でも動物病院などを舞台とした漫画を特定する概念アトリビュートとして利用可能である．C_M2 には役職を示す単語が集まっており，この概念アトリビュートが含まれる物語では病院内での政治闘争のような物語が期待できる．C_M3 や C_M4 ではサスペンスを連想させる事件に関連する単語が集まっている．一方で，C_M5 や C_M6 では，それぞれポジティブな感情と家族に関わる単語が集約されており，これらの概念アトリビュートが含まれる物語はハートフルな物語であることが想定でき，C_M3 や C_M4 の概念アトリビュートを有する物語とは性質が大きく異ることが予期される．また，C_M7 には病院内で処置される症状や処置に関わる単語が集約されており，具体的な医療行為が描かれる物語であることが示唆される．これらの概念アトリビュートは，2 章で示した基礎調査で見られた医療漫画というタグのみでは分類されていなかった詳細な内容を概念的に捉える上で有効活用できると考えられる．

4.3.3　ノイズと判定不可クラスタ

2 単語以上から構成されるクラスタについて，ファンタジー漫画では 3 クラスタ，医療漫画では 2 クラスがそれぞれノイズとなるクラスタであると評価できた．これらのクラスタには，品詞フィルタでは除去しきれなかった英数記号なで構成されていた．

多数の単語で構成されたクラスタについては，クラスタが表す概念が判定不可なものも存在した．ファンタジー漫画と医療漫画では表 6 と表 7 に示す概念判定不可クラスタが得られた．これらのクラスタ内には，メタ特性と内容特性，ノイズが混合している．クラスタ内を再クラスタリングすることで，一定の意味で整理された概念が理解可能になる可能性もある．精錬化の過程で対話的にクラスタの粒度を制御するしくみが必要である可能性が示唆される．

4.4　全体の考察

異なる 2 種類のジャンルの漫画のあらすじについて特性を分析したところ，あらすじで提示される情報が大きく異なることが確認できた．一般的な物語構造（起承転結など）に対応した内容のみが描かれるのではなく，ターゲット読者が求めているもの，あるいは，そのジャンルにおいて特徴化すべき情報に特化してあらすじが記述される傾向にあることが推察される．ジャンル間で共通する概念を有するかによって，既存の漫画ジャンル間での類似度を客観的に計算可能になる可能性もある．細かい単語の違いではなく概念レベルに量子化することで，人間の認知に親しいジャンル類似度が得られる可能性が考えられる．また，ジャンルごとではなく一般的な漫画全体を対象として概念アトリビュートを生成することで，既存のジャンルラベルとは違った観点で漫画の特性表現ができる可能性もある．

任意のドメインにおいて，あらすじ特性を評価のための妥当な概念アトリビュートを一度整備できれば，それ以降は整理された概念アトリビュートの集合を辞書として参照することで容易にあらすじを概念的に表現することが可能になる．このための概念アトリビュートの精錬化におけるインタラクションデザインも検討する必要がある．本稿での考察を通して，概念アトリビュートのク

ラスタリングの手法や粒度についてはさらなる議論が必要であることがわかった．また，クラスタに含まれる単語集合全体から想起される概念とは異なる概念の単語が少数紛れている事例も見られた．そのため，クラスタ内の単語集合の精錬とクラスタ自体の精錬に関するデザインについては今後の課題とする．

5 おわりに

本稿では，言語情報を対象として，コンテンツ特性の認識のための情報表現方法として概念アトリビュートを提案した．異なるジャンルの漫画のあらすじ集合に対して，概念アトリビュートを生成し，標本化された語彙による漫画のあらすじの特性表現の有用性について検討した．単純な方法である一方で，意味的に近い単語を概念としてまとめることで，人間の認識に親しい特性表現の実現可能性を示した．

今後は，提案手法での各処理手順におけるメタパラメータの設定方策や概念アトリビュート精錬化のインタラクションデザインについて検討する．また，漫画のあらすじ以外のドメインに対しても概念アトリビュートを生成し，コンテンツ特性の認識における有用性を議論していく．

謝辞

本研究では，eBookJapan に掲載された漫画あらすじを参照させていただいた．本研究は，立命館大学アートリサーチセンター日本文化資源デジタル・アーカイブ国際共同研究，科研費 JP20K12130，および，2021 年度国立情報学研究所公募型共同研究 (21S0501) の支援のもと行われた．記して謝意を表す．

参考文献

[1] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, *Advances in Neural Information Processing Systems 26*, pp. 3111–3119 (2013).

[2] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T.: Enriching Word Vectors with Subword Information, *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146 (2017).

[3] 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*, Minneapolis, Minnesota, Association for Computational Linguistics, pp. 4171–4186 (2019).

[4] 伊藤理紗，中村聡史： コミックにおける読者依存性の高い地雷表現の基礎調査とその軽減手法，第 5 回コミック工学研究会予稿集，pp. 18–25 (2021).

[5] 村井　源： 既存作品中の物語の基本パターンに基づく物語構造の自動生成，情報処理学会論文誌，Vol. 63, No. 2, pp. 335–346 (2022).

[6] Stevens, S. S., Volkman, J. and Newman, E. B.: A scale for the measurement of the psychological magnitude pitch, *Journal of the Acoustical Society of America*, Vol. 3, p. 185–190 (1937).

[7] Ellis, D. P. W. and Poliner, G. E.: Identifying 'Cover Songs' with Chroma Features and Dynamic Programming Beat Tracking, *Proceedings of 2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 4, pp. IV–1429–IV–1432 (2007).

[8] Xian, Y., Lampert, C. H., Schiele, B. and Akata, Z.: Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 9, pp. 2251–2265 (2019).

[9] 後藤和人，土屋誠司，渡部広一： 語彙の概念化と Wikipedia を用いた英字略語の意味推定方法，自然言語処理，Vol. 24, No. 3, pp. 351–369 (2017).

[10] 芦原和樹，梶原智之，荒瀬由紀，内田　諭： 多義語分散表現の文脈化，自然言語処理， Vol. 26, No. 4, pp. 689–710 (2019).

[11] Lowe, D. G.: Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110 (2004).

[12] Arthur, D. and Vassilvitskii, S.: K-means++: the advantages of careful seeding, *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms* (2007).

[13] Pelleg, D. and Moore, A.: X-means: Extending K-means with Efficient Estimation of the Number of Clusters, *Proceedings of the 17th International Conf. on Machine Learning*, pp. 727–734 (2000).

[14] Blei, D. M. and Jordan, M. I.: Variational inference for Dirichlet process mixtures, *Bayesian Analysis*, Vol. 1, No. 1, pp. 121 – 143 (2006).

[15] 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).

[16] 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, pp. 797–800 (2016).

[17] Miller, G. A.: The Magical Number Seven, Plus or Minus Two : Some Limits on Our Capacity for Processing Information, *The Psychological Review*, Vol. Vol.63, pp. 81–97 (1956).