Word2Vec を用いた卒業論文のクラスタリングと可視化

野田　拓哉（九州産業大学 理工学部 情報科学科）
Takuya NODA, Department of Information Science, Faculty of Science and Engineering, Kyushu Sangyo University
豊坂　祐樹（九州産業大学 産学共創・研究推進本部）
Yuki TOYOSAKA, Center of Industrial Collaboration, Kyushu Sangyo University
成　凱（九州産業大学 理工学部 情報科学科）
Kai CHENG, Department of Information Science, Faculty of Science and Engineering, Kyushu Sangyo University

1　はじめに

近年，学術情報のデジタル化が進み，膨大な数の論文が容易にアクセス可能となった．特に大学においては，毎年多数の卒業論文が作成されており，その内容を体系的に整理・把握することは，教育や研究の観点から極めて重要である [1][2]．しかし，卒業論文の多くはテキストベースで保存されており，内容の比較や分類を人手で行うには多大な労力と時間を要する．

このような課題に対して，自然言語処理技術の応用が注目されている [3][4]．中でも，Word2Vec は単語と文脈の関係をベクトルとして数値化する手法として知られ，文書間の意味的な類似度の定量化が可能である．これにより，論文内容の特徴を抽出し，意味的に近い論文同士を自動的にクラスタリングする手法の開発が期待される．

一方で，Word2Vec は個々の単語に焦点を当てた手法であるため，文書全体の特徴を適切に捉えるためには前処理や文ベクトルへの変換が不可欠となる．また，クラスタリング結果をどのように可視化し，ユーザーにとって直感的に理解可能な形で提示するかも重要な課題である．特に，クラスタの妥当性や論文の主題間の関係性を示すには工夫が求められる．

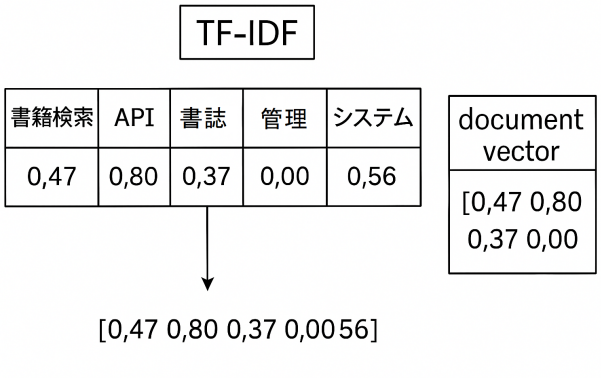


図 1: TF-IDF 重みによる文書ベクトル例

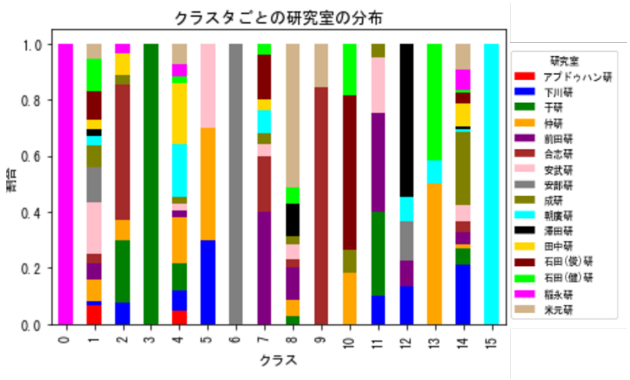


図 2: TF-IDF に基づく文書クラスタリング

本研究では，日本語の卒業論文題目を対象に，形態素解析と Word2Vec を組み合わせたテキストクラスタリング手法を用いて，研究分野のトレンドやテーマ間の関係性を可視化する．これにより，多数の研究題目がどのような分野に分類され，どのような関連性を持つかを直感的に理解できる可視化手法を提案する．これにより，研究動向の把握，類似研究の発見，さらには学生間の知的交流の促進など，大学教育や研究支援に資する基盤的手法の構築を目指す．

具体的には，まず，単語間の意味的な類似性を考慮できるベクトル化手法 Word2Vec を導入することで，同義語や類似語を適切に処理することが可能となる．また，クラスタリング結果を視覚的に示し，中身の比較と考察を行うことで，提出手法の有効性を評価する．

2　TF-IDF に基づく文書クラスタリング

本研究室では，TF-IDF（Term Frequency-Inverse Document Frequency）の重み付けを用いてテキストデータをベクトル化し，クラスタリングを実施してきた [5][6][7]．TF-IDF は文書に含まれる単語の重要度を評価する手法である．ある文書の中で，特定の単語がどれだけ頻繁に出現するか（TF）と，その単語が他の文書ではどれだけ少ない頻度で出現するか（IDF）を掛け合わせることで，

単語の重要度を数値化する．図 1 は TF-IDF に基づいて文書をベクトルとして表現する仕組みを示している．

各単語の TF-IDF スコアを並べて 1 つのベクトル（数値の配列）にすることで，文書全体を数値で表現できる．このベクトル表現は，文書の類似度計算やクラスタリング，分類タスクなどに利用される．先行研究 [5][7] では，卒業論文題目を対象としてクラスタリングを行い，類似研究を行う研究室グループを分析することを試み，図 2 のような結果が得られた．

しかし，この方法は，単語の出現頻度と重要度に基づいてテキストデータを数値化するものであり，いくつかの課題が存在し，十分な成果が得られなかった．まず，意味の近い単語でも表現の違いによって別の単語として扱われるという問題があった．たとえば，「ウェブ」と「Web」など，同義語や派生語が異なる単語として認識されるため，同じクラスタに分類されないケースが発生していた．これは，TF-IDF が単語単位での重み付けに依存しているため，単語間の意味的な関連性を十分に捉えられないことに起因している．

また，クラスタリング結果の評価が適切に行われていなかった．クラスタリング性能を定量的に評価する指標を用いず，主に主観的な判断に基づいて評価していたため，結果の信頼性や妥当性が十分に保証されていなかった．

さらに，クラスタリング結果の可視化が行われていなかったことが挙げられる．可視化が欠如していたことで，各クラスタに含まれる要素やクラスタ間の関係性が把握しづらく，研究結果の解釈が難しい状態であった．可視化はクラスタリング結果の透明性や直感的理解を促進する重要な手段であるが，その必要性が十分に考慮されていなかった．

3 単語の分散表現 Word2Vec

Word2Vec は，Google によって開発された自然言語処理のためのアルゴリズムであり，単語を高次元ベクトル空間に埋め込む技術である．この技術の特徴は，単語間の意味的な類似度を数値的に計算できる点にあり，例えば「王」＋「女性」－「男性」＝「女王」のような意味的な関係をベクトル演算によって導き出すことが可能である．Word2Vec は，膨大なテキストデータから単語の埋め込みベクトルを学習し，その結果として得られる単語ベクトルは，機械学習の様々なタスクで有用な特徴量となる．

Word2Vec では，CBoW（continuous bag-of-words）モデルおよび skip gram モデルという二つのモデル構造のいずれかを使用し，単語の分散表現を生成する [8][9]．**CBOW（Continuous Bag of Words）**は文脈（周囲の単語）からターゲットとなる単語を予測するモデルである．高速に学習でき，一般的な用途に適している．CBOW は次の損失関数を用いてモデルを訓練する．

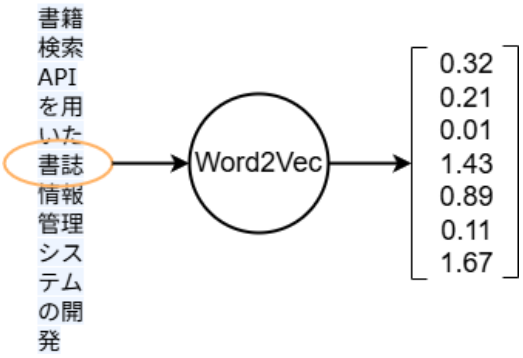


図 3: Word2Vec による「書誌」の単語分散表現

$$-\frac{1}{T}\sum_{t=1}^{T}\log P(w_t|w_{t-1},w_{t+1}) \quad (1)$$

ここは T: コーパス内の単語数, w_t: ターゲットの単語である．この損失関数をできるだけ小さくすることで，CBOW モデルの学習を行うことができる．

Skip-gram は，ターゲット単語から文脈を予測するモデルである．CBOW に比べて計算コストが高いが，希少な単語の学習に強い．Skip-gram では次の損失関数を用いてモデルを訓練する．

$$-\frac{1}{T}\sum_{t=1}^{T}(\log P(w_{t-1}|w_t)+\log P(w_{t+1}|w_t)) \quad (2)$$

Skip-gram モデルは，コンテキストの数だけ推測を行うため，損失関数は各コンテキストで求めた損失の総和を求める必要がある．一方，CBOW は一つのターゲットの損失を求める．

TF-IDF に比べると，TF-IDF は一般的な単語の影響を減少させ，文書における重要な単語を強調するが，数万次元の単語ベクトルに多くの単語が 0 に近い値を持つため，データがスパース（疎）になる．一方，Word2Vec は数百次元となる密な特徴量を生成し，意味的に関連する単語を近くに配置することで，単語の意味をより良く捉えるため，大規模なデータセットに対して効率的に学習できる．

4 文書クラスタリングについて

クラスタリングとは，ある特徴量空間上のデータを複数個のクラスタに分類する手法である．クラスタリングは「教師なし学習」の 1 種であり，「クラスタ解析」や「クラスタ分析」などと呼ばれることもある．クラスタリングにはハードクラスタリングとソフトクラスタリングの 2 種類がある．各デー

タが単一のグループに所属するようにグルーピングする手法をハードクラスタリングと呼び，各データが複数のグループに所属することを許容してグルーピングする手法をソフトクラスタリングと呼ぶのである．

クラスタリングの計算方法は大きく「階層クラスタリング」と「非階層クラスタリング」の 2 つに分けられる [10][11]．階層クラスタリングでは階層構造を構築しながらグループ化を進めるのに対し，非階層クラスタリングは階層構造を作らずに分類を行う．それぞれの手法には特徴があり，データの量や分析目的に応じて使い分けることが重要である．

非階層クラスタリングとは，階層を作らずクラスタ数を事前に決め分類していく手法である．非階層クラスタリングは，階層クラスタリングと違い事前にクラスタ数を決めて指定しておく必要がある．あらかじめ決められたクラスタ数に要素を分類していくため，データ同士の距離を計算する階層クラスタリングより計算量を少なくすることができる．そのため，データ量の多いデータの分析に適した手法である．

階層クラスタリングと非階層クラスタリングの違いは，サンプルサイズの多さに依存する．サンプルサイズが 100 以下の比較的小さい場合には階層クラスタリングが使用される．それ以上のデータサイズは非階層クラスタリングが使われる場合が多い．

4.1 文書クラスタリングの手順

文書集合に対してクラスタリングを適用する手順について述べる．

a. クラスタリングの対象を決める

クラスタリングの対象として「サンプル（データ点）」を選ぶ場合は，各サンプルをグループ化することが目的である．例えば，顧客のクラスタリングでは顧客ごとにグループを作成する．逆に「変数」をクラスタリングする場合は，データの次元削減や特徴の抽出を目的とする．例えば，複数の変数をグループ化し，それらを 1 つの代表的な変数にまとめることができる．対象を選ぶ際は，目的に応じて適切なアプローチを選ぶことが重要である．

b. クラスタリングの手法を決める

クラスタリング手法を選択する際には，データの特性や目的に合わせた手法を選ぶことが必要である．特に，データ数が多い場合やグループの過程が重要かどうかを考慮する．群平均法やウォード法は，データ数が少なく，グループがどのように分かれていくかの過程を知りたい場合に使用される．これらは階層的クラスタリングにおいてよく使用される手法である．k-means 法は，データ数が多く，グループ数があらかじめ決まっている場合に有効である．k-means 法は，大規模なデータセットに対しても高速に実行でき，効率的である．これらの手法の選択は，最終的にどのような分析結果を求めるかによって決まる．

c. データ間の類似度の尺度を決める

クラスタリングでは，データ間の類似度を測るために距離を使用する．距離はデータ点同士の近さを示し，一般的に距離が短いほど似ているとされる．代表的な距離尺度には以下のようなものがある．

- **ユークリッド距離**: 空間的な距離を基準にしたもので，最も一般的に使用される．数値データに適している．
- **マンハッタン距離**: 各軸方向に沿った距離の合計を基に計算する距離で，異なるタイプのデータにも使用できる．
- **コサイン類似度**: ベクトルの方向の類似性を測るため，主にテキストデータや高次元データに使用される．

データの性質に応じて適切な距離の種類を選ぶ必要があり，例えばテキストデータではコサイン類似度を使用することが多い．

4.2 k-means 法

クラスタの数（k 個）を決め，ランダムに選んだデータをクラスタ中心とし，残りのデータを最も近いクラスタ中心に割り当てることでクラスタを形成する手法である．以下の流れでクラスタリングを行う．

1. 代表点をランダムに決める．この際にクラスタが 3 つなら 3 つの代表点ができる．
2. クラスタに分ける．ユークリッド距離を計算して各データがどの代表点に近いかでグループ分けを行う．
3. 重心点を計算する．このクラスタの重心点が次の代表点になる．
4. クラスタの再分類

k-means 法は，計算がシンプルで実行速度が速く，大規模データにも対応可能なアルゴリズムである．一方で，クラスタ数を事前に指定する必要があることや，初期値の設定によって結果が異なる可能性がある点が課題である．アルゴリズム名の「k」はクラスタ数を示すハイパーパラメータであり，この値はデータ数よりも小さい必要がある．適切なクラスタ数を選ぶことが結果の精度に影響を与える．

5 卒業論文のクラスタリング

卒業論文クラスタリングの実験手順として，まず，卒業論文題目をテキストデータとして収集する．次に収集したテキストデータに対して前処理を行い，Word2Vec を用いて文書ベクトルを計算する．文書ベクトルをクラスタリングする．最後にクラスタリング結果を散布図，積み上げグラフ，ワードクラウド等を利用して可視化する [12]．

5.1 テキストデータの前処理の流れ

本研究で行った前処理として，1. 形態素解析の実施，2. ストップワードの摘出と名詞の抽出，3. Word2Vec モデルの訓練，4. 文ごとの平均ベクト

ルの計算を行う．

形態素解析とは，自然言語（日本語や英語など）の文章を，意味を持つ最小単位である「形態素」に分割し，それぞれの形態素がどのような品詞（名詞，動詞，助詞など）に属するのかを識別する処理 [13] である．日本語形態素解析に，MeCab, Janome 等のツールがよく用いられる．本研究では，形態素解析後にストップワードの削除と名詞の抽出を行う．

5.2 文の分散表現を求める

本研究では，文の意味を数値的に表現するために，Word2Vec を用いて文ごとの分散表現を求める手法を採用した．Word2Vec を利用することで，対象の意味の類似性を捉えることができる．これにより，単語間の意味的な関係を数値的に捉えることが可能となり，文全体の意味を表現するために有効である．

・形態素解析による名詞抽出

最初に，対象となる文を形態素解析ツールである MeCab を使用して処理した．MeCab を使用することで，文を単語単位に分割し，各単語の品詞を得ることができる．去年度とできるだけ条件を統一させるため，文の主な意味を担っている名詞のみを抽出した．また，無意味な単語である「ため」をストップワードとして除外した．

この処理により，各文から名詞のみが抽出され，その後，Word2Vec モデルに入力される．

・Word2Vec モデルの学習

次に，形態素解析により得られた名詞のリストを基に，Word2Vec モデルを学習させる．Word2Vec は，指定したウィンドウサイズ内で単語の関連性を学習し，単語同士の意味的な関係をベクトルで表現する．今回使用したモデルは，200 次元のベクトル空間を採用し，最小頻度 1 の単語も含めて学習を行う．これにより，文の中の単語を適切にベクトル化することができる．

・題目のベクトル化

単語の埋め込みベクトルを用いて文の埋め込みベクトルを求める基本的な方法はいくつか存在するが, 代表的な手法のひとつは**平均プーリング（mean pooling）**である．平均プーリングによる文ベクトルは次のように求める．

文 S が n 個の単語 $w_1, w_2, \ldots, w_n$ から構成され，それぞれの単語に対応する埋め込みベクトルが $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n$ であるとき，文の埋め込みベクトル $\boldsymbol{s}$ は次式で表される：

$$\boldsymbol{s} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{w}_i \tag{3}$$

平均プーリング法で各題目の意味を表現するために，題目の単語ベクトルの平均を取り，題目のベクトルを計算する．具体的には，各題目に含まれる単語ベクトルを抽出し，その平均を取ることで題目全体を代表するベクトルを求める．これにより，題目全体の意味を数値的に表現することができる．図 4 は平均プーリングによる題目の埋め込み表現を計算するイメージを示す．

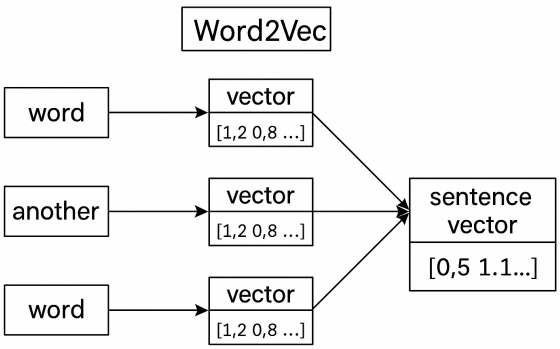


図 4: 平均プーリングによる題目分散表現計算

5.3 クラスタリングの実行

本研究では，九州産業大学理工学部情報科学科の令和 1 年度から 4 年度までの全ての卒業論文のタイトル 491 個をデータとしてクラスタリングを行う．去年度と同じ条件下にするため k-means 法の k の値を 16 としてクラスタ 0 からクラスタ 15 にグループ分けする．また，様々な考察を立てるため k の値を 10，25 といったグループ分けも行う．

6 実験結果

この章では，学位論文テキストデータを利用し，前処理を行ったテキストデータを Word2Vec を使用して，各題目を構成する単語のベクトルを学習し，文ごとの単語ベクトルの平均を計算して文全体のベクトルを求める．これらのベクトルは，後でクラスタリングの際に使用される．

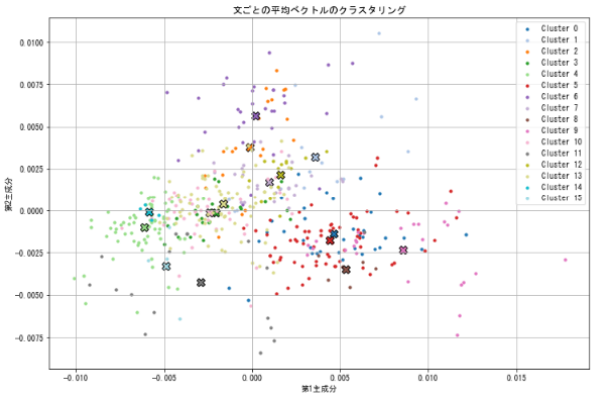


図 5: クラスタリング結果（$k = 16$）の散布図

6.1 クラスタリングの結果

図 5 は k の値を 16 にした場合の散布図を示す．ここでは，200 次元あるベクトルを 2 次元削減している．横軸 (第 1 主成分) は，PCA（主成分分析）によって求められた最も重要な軸で，データの分散が最も大きい方向に沿った成分である．この軸は，文ごとのベクトルの中で最も重要な特徴を捉えている．この軸に沿って文の特徴が分布している．縦軸（第 2 主成分）は，第 1 主成分の次にデータの分散が大きい方向を示す軸である．この軸は横軸と直交しており，主に横軸で捉えきれなかった特徴を表現する．

k-means 法によって得られた各クラスタの最も中心に近い題目について考察する．k-means 法における重心は，クラスタ内の平均的な位置を表す点である．下記に各クラスタの重心に最も近い場所にある卒業論文題目を一部列挙する．

公共交通基盤データを用いた路線図表示Webアプリケーションの改良	稲永研	R03	Cluster 0
公共交通基盤データ作成支援Webアプリケーションにおける運行日情報作成機能の試作	稲永研	R03	Cluster 0
公共交通基盤データを用いた車内案内表示Webシステムにおける天候情報表示機能の追	稲永研	R03	Cluster 0
プライバシー保護データ解析における匿名化処理およびデータ拡張	成研	R03	Cluster 0
地理空間情報システムを使ったデータ視覚化手法調査及び地域公共交通運行管理支援シ	稲永研	R04	Cluster 0
GTFSデータ作成支援ツールの調査	安武研	R04	Cluster 0
デマンド交通運行管理システムにおける運行実績データ表示機能の開発	稲永研	R04	Cluster 0
LSTMモデルを用いた時系列データ分類に関する演習開発	前田研	R04	Cluster 0
プライバシー保護データ解析のためのデータ拡張	成研	R04	Cluster 0
公共交通基盤データを用いた車両運行履歴表示Webシステムの開発	稲永研	R04	Cluster 0

図 6: クラスタ 0 に含まれるテーマ例

クラスタ 0 で最も重心に近い題目は，稲永研が行っている「Web 地図表示ライブラリを用いた地域公共交通向けアンケート調査データの視覚化システムの開発」である．クラスタ 0 内には，稲永研，合志研，成研，安武研，前田研，于研が含まれ，データに関する研究が分類されている．

リプレイを用いた車間距離維持教育用のドライビングシミュレータの開発	合志研	R01	Cluster 1
ドライビングシミュレータにおける踏切の開発	合志研	R01	Cluster 1
あおり運転対策ドライビングシミュレータの開発	米元研	R02	Cluster 1
Webブラウザ上で動作する住宅街における安全運転教育用ドライビングシミュ	合志研	R02	Cluster 1
ドライビングシミュレータにおける方向指示器機能の実装のための教材の作成	合志研	R03	Cluster 1
ドライビングシミュレータ開発における複数車両の走行に関するプログラミン	合志研	R03	Cluster 1
一時停止及び進入禁止についての学習用ドライビングシミュレータの開発	合志研	R03	Cluster 1
信号停止による練習走行機能を持つ車間距離維持教育用ドライビングシミュレ	合志研	R03	Cluster 1
追い越し可否についての学習用ドライビングシミュレータの開発	合志研	R03	Cluster 1
中型車両の安全運転教育用ドライビングシミュレータの開発	合志研	R04	Cluster 1
ドライビングシミュレータにおける他車開発のためのUnity ML-Agentsについ	合志研	R04	Cluster 1
車間距離維持教育用ドライビングシミュレータの改良	合志研	R04	Cluster 1
PLATEAUを用いたドライビングシミュレータのための実車走行データに基づく	合志研	R04	Cluster 1
PLATEAUを用いたドライビングシミュレータのための道路生成機能の改良	合志研	R04	Cluster 1

図 7: クラスタ 1 に含まれるテーマ例

クラスタ 1 で最も重心に近い題目は，合志研が行っている「Web ブラウザ上で動作する住宅街における安全運転教育用ドライビングシミュレータの開発」である．クラスタ 1 内には，合志研が行っている「ドライビングシミュレータの開発」，米元研が行っている「VR 事故対策シミュレータの開発」など，運転や交通事故に関連するシミュレータの開発に関する研究が分類されている．

このクラスタでは，運転に関わる卒業論文題目のみが存在する．単語の分散表現を求めた際，名詞である「運転」や「交通」といった単語がベクトル空間上で近い位置にあり，文の分散表現を求めた際に似た題目として分類されたと考える．

自然数の和問題に対するアルゴリズムの実装と実験による性能評価	朝廣研	R01	Cluster 12
Java言語によるアルゴリズムの実装と評価	朝廣研	R01	Cluster 12
アルゴリズムの実装と実験による 性能評価 ―自然数の和問題とべき乗問題―	朝廣研	R02	Cluster 12
アルゴリズムの実装と実験による 性能評価 ―素数判定問題に対して―	朝廣研	R02	Cluster 12
サーバの可用性を高めるための性能評価と監視	下川研	R02	Cluster 12
アルゴリズム実装と実験による性能評価 ―多項式問題―	朝廣研	R03	Cluster 12
素数判定問題に対する試し割り法の 平方根を用いた改良とその性能評価	朝廣研	R03	Cluster 12
バブルソートの処理時間に関する実験と評価	朝廣研	R03	Cluster 12
約数の個数問題に対するアルゴリズムの性能評価 ―2018年度貸与PCを用いた実験―	朝廣研	R04	Cluster 12
入れ子型和問題に対するアルゴリズムの性能評価 ―2019年度貸与PCを用いた実験―	朝廣研	R04	Cluster 12
自然数の和問題に対するアルゴリズムの性能評価 ―2019年度貸与PCを用いた実験―	朝廣研	R04	Cluster 12

図 8: クラスタ 12 に含まれるテーマ例

クラスタ 12 で最も重心に近い題目は，「入れ子型和問題に対するアルゴリズムの性能評価 ―2019 年度貸与 PC を用いた実験―」である．クラスタ 12 内には，下川研と朝廣研が行っている性能評価が分類されている．

また，このクラスタには性能評価に関連する卒業論文題目のみが存在しており，文の分散表現を求めた際に「性能」や「評価」という単語を共有していることから，同じクラスタに分類されたと考えられる．

酵素反応ネットワーク分割表現による適応応答系の探索	仲研	R03	Cluster 14
酵素反応ネットワーク分割表現による振動応答系の探索	仲研	R03	Cluster 14
振動応答系の酵素反応ネットワークにおけるロバスト性の解析	仲研	R04	Cluster 14
適応応答系の酵素反応ネットワークにおけるロバスト性の解析	仲研	R04	Cluster 14

図 9: クラスタ 14 に含まれるテーマ例

クラスタ 14 で最も重心に近い題目は，仲研が独自に行っている「酵素反応ネットワーク分割表現による適応応答系の探索」である．クラスタ 14 内には，仲研の「酵素反応ネットワーク」に関わる研究のみが存在する．

このクラスタでは，酵素反応ネットワークに関連する卒業論文題目のみが存在している．これは「酵素」，「反応」，「ネットワーク」という単語が文の分散表現を作る際に，影響を及ぼしており，独自のクラスタとして現れたと考えられる．

図 2 は k の値を 16 にした場合に，昨年度のクラスタリング結果の積み立て棒グラフを示しており，図 10 は，Word2Vec に基づく文書ベクトルを用いたクラスタリング結果の積み上げ棒グラフを示している．

図 2 と図 10 を比較すると，クラスタ 0 内が本研究ではデータに関する研究が含まれていたが昨年度では稲永研の研究のみが分類されている．クラスタ 12 では，性能評価の実験が行われている．昨年度は，米元研，石田（健）研，合志研，仲研といった性能評価とは関係ない実験が行われている研究室が分類されていた．

これらのことから，本研究のクラスタリング手法は，研究題目の内容をより正確に反映した分類が可能であることが示唆される．具体的には，クラスタ

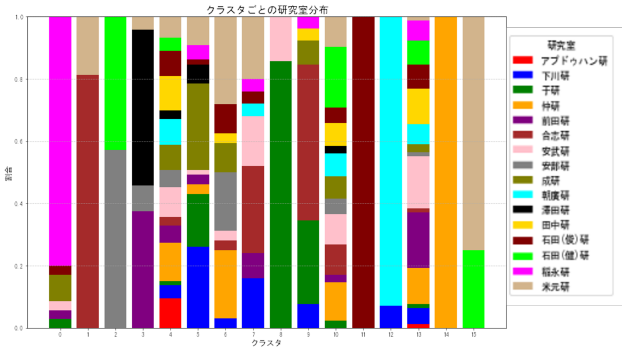


図 10: Word2Vec 結果の積み上げグラフ（$k = 16$）

0 においてデータに関する研究が適切に分類されている点や，クラスタ 12 において性能評価に関連する研究が集約されている点から，研究テーマの類似性を捉える精度が向上していると考えられる．

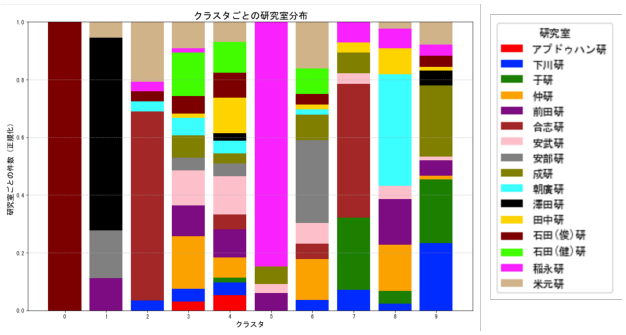


図 11: Word2Vec 結果の積み上げグラフ（$k = 10$）

図 11 に k の値を 10 にした場合のクラスタリングの積み上げグラフの結果を示す．$k = 10$ に設定した場合，独自の研究を行っている石田（俊）研が単独のクラスタとして出現した．石田（俊）研では，「セルオートマトンを用いたストリーム暗号システムおよび可逆性について」など，他の研究室では扱われていないテーマの研究が数多く行われている．このため，石田（俊）研が独立したクラスタとして現れたと考えられる．

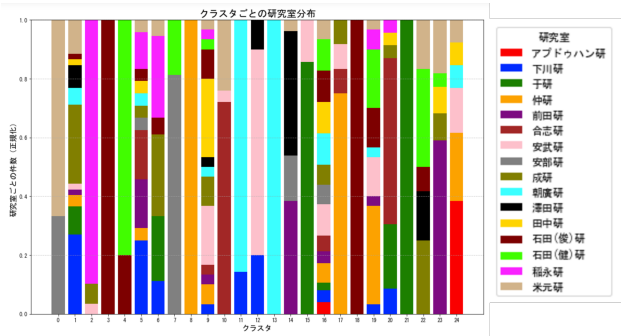


図 12: Word2Vec 結果の積み上げグラフ（$k = 25$）

図 12 に k の値を 25 に変更した場合のクラスタリングの積み上げグラフの結果を示す．$k = 25$ に設定すると，より多様なジャンルの研究に分けられるため，$k = 10$ の結果と比較してクラスタごとの独自性が強まっている．例えば，クラスタ 13 に分類される朝廣研究室では，「約数の個数問題に対するアルゴリズムの性能評価 —2018 年度貸与 PC を用いた実験—」といった独自の性能評価実験に関する研究が含まれている．また，クラスタ 8 には，仲研究室による「酵素反応ネットワーク分割表現による適応応答系の探索」など，酵素反応ネットワークに特化した研究のみが含まれている．

これらのことから，k の値を小さく設定すると，似た研究が同じクラスタに含まれやすい．しかし，卒業論文題目は異なっていても，ベクトルの数値が近いため，同じクラスタ内に含まれることがある．一方で，k の値を大きくすることで，ジャンルごとにより細かく分類され，各クラスタ内のデータがより類似したものとなり，特徴が k の値を小さくした場合よりも具体的に卒業論文題目が反映される．

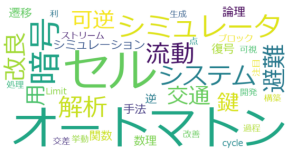
6.2 クラスタリング結果の可視化

ワードクラウド（Word Cloud）とは，テキストデータ内で頻繁に登場する単語を，その頻度に応じて文字の大きさを変えて視覚的に表現する手法である [14]．ワードクラウドを使うことにより，テキストデータに隠された傾向や重要なキーワードを直感的に把握することができる．

図 13〜図 15 は $k = 10, 16, 25$ のクラスタリング結果をワードクラウドを用いて可視化したものを示す．「セルオートマトン」といった石田 (俊) 研独自の研究が，k の値に依存せず独立したクラスタを形成した．しかし，川研，于研，合志研，成研，仲研，稲永研が行っていた「管理」に関する研究は，k の値を小さくすることで一つのクラスタに分類された．一方で，k の値を 25 のように大きくすることで，交通事故を防止するための管理に関する研究と地域公共に向けた管理に関する研究に分かれた．

これらの結果から，k の値を増やすことでより卒業論文題目が詳しく分類されたことがわかった．一方で，k の値を減らすことで大きな枠組みの中で分類される．これらのことから，k の値の選択は，研究テーマの細分化や統合の度合いに大きな影響を与えることがわかる．そして，テーマごとの関係性や研究の特徴をより詳細に理解するためには，k の値を調整することでジャンルを細分化したり，大まかなジャンル分けを行ったりすることが有効であると考えられる．

特に，k を小さく設定することで，関連性の高い研究テーマが一つのクラスタに統合され，共通性を強調することが可能となる．一方で，k を大きく設定することで，テーマ間の違いや特異性を反映した細分化が可能となり，研究分野の多様性や特徴を浮き彫りにすることができる．



(a) クラスタ 0

(b) クラスタ 3

(c) クラスタ 5

(d) クラスタ 7

図 13: $k = 10$ のワードクラウド

(a) クラスタ 2

(b) クラスタ 3

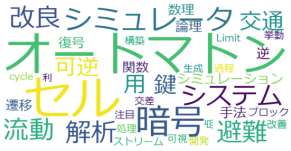
(c) クラスタ 10

(d) クラスタ 18

図 15: $k = 25$ のワードクラウド

(a) クラスタ 1

(b) クラスタ 9



(c) クラスタ 11

(d) クラスタ 14

図 14: $k = 16$ のワードクラウド

7 おわりに

本研究は，日本語の卒業論文題目を対象に，形態素解析と Word2Vec を組み合わせたクラスタリング手法を用いることで，研究分野のトレンドやテーマ間の関係性を可視化することを目的としている．また，昨年度の研究では，TF-IDF を用いてテキストデータを数値化し，クラスタリングを実施したが，いくつかの課題が明らかになった．

まず，TF-IDF では単語の出現頻度と重要度に基づいてデータを数値化できるという利点があるものの，意味の近い単語が表記の違いによって別の単語として扱われるという問題があった．たとえば，「ウェブ」と「Web」などの同義語や派生語が異なる単語として認識されるため，同じクラスタに分類されないケースが発生した．これは，TF-IDF が単語単位での重み付けに依存しており，単語間の意味的な関連性を十分に捉えられないことに起因している．また，クラスタリング結果の評価においても，性能を定量的に評価する指標が用いられず，主に主観的な判断に基づいて評価されていたため，結果の信頼性や妥当性が十分に保証されていなかった．

これらの課題を解決するため，本研究では，形態素解析による名詞抽出とストップワードの除去を行い，その後，Word2Vec を用いて題目を分散表現に変換する手法を採用した．Word2Vec は，単語間の意味的な類似性を考慮することで，従来の TF-IDF の限界を克服し，表記ゆれや同義語を適切に処理できる特徴がある．題目ごとの分散表現を計算した後，k-means 法を用いてクラスタリングを実施し，各クラスタにおける代表的な題目を抽出することで，分野の傾向やテーマの関連性を視覚的に把握可能にした．

さらに，クラスタリング結果をワードクラウドや散布図などの可視化手法を用いて表現し，クラスタ間の関係性や特徴を明確化した．この可視化により，膨大な研究題目を効率的に分類するだけでなく，各クラスタ内の研究内容の概要を直感的に把握できる点が特徴である．本研究は，特定の研究室や分野におけるテーマの偏りを分析し，研究トレンドを把握するための有用なツールとして応用が期待される．

今後は，クラスタリング手法の改良や，他の分散表現モデルとの比較検討を進め，さらなる精度向上を目指す．また，クラスタリング結果や可視化結果をどのように有効利用するかについても検討を行う予定である．

謝辞

本研究を進めるにあたり，成研究室所属大学院修士 2 年の井手先輩にお忙しい中，様々なご指導ご鞭撻をいただきました．心から感謝いたします．また，本研究で使わせていただいた学位論文等検索ホームページを作ってくださった方にも心から感謝いたします．

参考文献

[1] C. Lee Giles (2013). Scholarly big data: information extraction and data mining. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management (CIKM '13). Association for Computing Machinery, New York, NY, USA, 1 – 2. https://doi.org/10.1145/2505515.2527109

[2] F. Xia, W. Wang, T. M. Bekele and H. Liu (2017). Big Scholarly Data: A Survey, in IEEE Transactions on Big Data, vol. 3, no. 1, pp. 18-35, 1 March 2017, doi:10.1109/TBDATA.2016.2641460.

[3] 張馨雲, 今泉優気, 隈部晶, 林成元, 豊坂祐樹, 成凱, テキスト解析及び機械学習による卒業研究テーマトレンドの可視化, 火の国情報シンポジウム 2023, 2023 年 3 月.

[4] 張馨雲 (2024), 学術情報のテキスト解析による研究動向の可視化, 九州産業大学理工学部情報科学科卒業論文, 2024 年 1 月.

[5] 隈部晶 (2023), 卒業論文テーマのクラスタリングによる研究室配属支援, 九州産業大学理工学部情報科学科卒業論文, 2023 年 1 月.

[6] 今泉優気 (2023), テキスト解析による卒業研究テーマの可視化, 九州産業大学理工学部情報科学科卒業論文, 2023 年 1 月.

[7] 織方鵬宇 (2024), 卒業論文の統計解析による研究室配属支援, 九州産業大学理工学部情報科学科卒業論文, 2024 年 1 月.

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13), Vol. 2. Curran Associates Inc., Red Hook, NY, USA, 3111 – 3119.

[9] T. Mikolov, K. Chen, G. Corrado and J. Dean (2013). Efficient Estimation of Word Representations in Vector Space. 2013. arXiv:1301.3781

[10] 富田浩世 (2019), 最初に学ぶクラスタリングの特徴と種類, 株式会社 AUC 技術ブログ, 2019.8.6. https://www.acceluniverse.com/blog/developers/2019/08/post-17.html (2024 年 12 月 3 日閲覧)

[11] こがたぶろぐ【機械学習】階層化クラスタリングの 6 つの手法 | グループ分類に必要な知識, こがたブログ 2020.12.1. https://kgt-blog.com/tech-17/2264/ (2024 年 12 月 9 日閲覧)

[12] 三末和男 (2021), 情報可視化入門 : 人の視覚とデータの表現方法, 森北出版 (2021/6/1)

[13] 工藤拓 (2018), 形態素解析の理論と実装, 株式会社近代科学社, (2018)

[14] Andreas Darmawan, シンプルかつ直感的ー自然言語処理における Wordcloud の活用. A&P Global Blog 2021.3.11. https://www.spglobal.com/marketintelligence/jp/news-insights/blog/message-in-a-word-cloud (2024 年 12 月 9 日閲覧)