学術情報のテキスト解析による生体医工学研究動向の可視化

張　馨雲[1]　林　成元[1]　成　凱[1,a)]

概要：学術情報の急速な電子化と大規模化に伴い，データ解析・可視化を通して専門分野の動向を把握・予測し，新しい課題を開拓することが期待されている．しかし，専門性の高い学術情報を解析する際には専門用語抽出等，様々な課題が存在する．本研究では，生体医工学を対象としてデータ収集，テキスト解析を行い，学術情報の可視化を試みる．主要なフローとしては，データ収集，データ前処理，テキスト解析，可視化で構成されている．本研究の結果では，生体医工学の研究動向を包括的に把握し，ワードクラウドや ThemeRiver などの可視化手法を活用してデータを効果的に提示した．これにより，研究者や研究トピックの重要性が視覚的に理解でき，今後の方向性も把握できる．今後の課題として，より高度で効果的なデータ解析手法や異なる可視化手法の組み合わせによる情報理解の向上が挙げられ，データ収集の最適な方法も検討される予定である．

Visualization of Biomedical Engineering Scholarly Data by Text Analysis

1. はじめに

近年，学術情報の電子化・大規模化が急速に進展し，これに伴い学術データの分析のニーズが飛躍的に増加している．研究者たちはこれらのデータを通して新たな洞察を得ると同時に，その成果を効果的に共有する手段として「データ可視化」に注目を浴びている [1][2]．データ可視化（Data visualization）とは，数値情報だけでは理解しづらい複雑な現象や関係性，変化を，わかりやすい形（例えば，グラフ，チャート，表，画像など）に変換するプロセスである．これにより，データの中に潜む情報を視覚的に提示し，数字から得る情報の理解をサポートする．この可視化はしばしば「見える化」や「視覚化」とも呼ばれ，大規模なデータセットに潜むパターンや傾向，関係性，外れ値などを容易に識別できる利点がある．

しかし，「可視化」には一概に適用できる決まりごとは存在せず，分野や解析目的，データセットに応じて適切なデータ処理方法や可視化方式を選定することは容易ではない．なぜなら，データには数値だけでなく，日付，テキスト，カテゴリなどが含まれ，その性質によって最適な可視化手法が異なるからである．また，データが単なる静止状態を表すだけでなく，時間に伴う変化や傾向を可視化することも重要である．大規模で複雑なデータを扱う際には，可視化の対象を適切に決定することも難しくなる．さらに，専門性の高い学術情報を解析する際にはその分野の特有な専門用語やトピックを効率よく抽出する等様々な課題が存在し，データ解析と可視化を適切に行えるとはいえない状況である [2]．

本研究は，生体医工学分野の学術情報（とりわけ学術論文の題目）を対象とし，学術情報の解析と可視化を通して研究動向の把握を目指している．その中で，過去数年間のトピック，研究テーマ，新たな手法，応用分野などを具体的に掘り下げ，最新かつ包括的な情報を獲得することが主要な目的である．まず，研究トレンドの特定では，テキスト解析手法を通じて最新の生体医工学研究におけるトレンドや熱点を明確に把握する．これにより，研究者が注目すべき重要なキーワードやトピックを把握し，研究の最前線にどのように関与できるかを理解する．また，テキスト解析手法の有用性の検証を行う．特に脳波解析分野においてこの手法が有益であるかを示し，研究動向の可視化が研究者や学術コミュニティに与える具体的な付加価値を明確にしたい．

[1] 九州産業大学
Kyushu Sangyo University,
2-3-1, Higashi-ku, Fukuoka, 813-8503, Japan
[a)] chengk@is.kyusan-u.ac.jp

2. データ可視化の基本事項

本節では，テキスト解析と可視化の基本事項をまとめる．まず，情報可視化を実施する際に最も重要なのは，目的を明確に定義することである．可視化の目的によって，使用する手法やチャートの選択が異なる．データから何を伝えたいのか，何を知りたいのかを事前に把握することが，効果的な可視化の出発点となる．

データの性質も考慮し，数量データ（数値），質的データ（カテゴリ），時系列データなど，データの種類によっては異なる可視化手法が必要である．データの性質を理解し，それに合わせた適切な可視化手法を選ぶことが求められる．

可視化手法の種類も多岐にわたり，棒グラフ，折れ線グラフ，散布図，ヒートマップ，パイチャートなどがある．これらはそれぞれ異なるデータの特性や伝えたい情報に適している．また，適切なカラーパレットやコントラストの選択も注意が必要である．色は情報の視覚的な強調や区別に役立つ一方で，濫用は避けるべきだと思われる．

さらに，可視化には物語を組み込むことで，専門家や部外者などがデータの意味を理解しやすくなる．順序立てて情報を提示し，データにストーリーを与えることが効果的である．また，データの信頼性は可視化の基盤となる．正確な計測とデータのクリーニングが行われていることを確認することで，信頼性の高い可視化が可能である．

最後に，データの性質や目的に応じて適切な可視化ツールを選択することが必要となり，柔軟性と多様性を兼ね備えているツールを選択すべきである．情報可視化はデータの理解と意思決定をサポートする力強な手法であり，これらの基本事項を遵守することで，洞察に富んだ効果的な可視化が可能となることを期待される．

本研究の対象となった生体医工学分野の学術論文題目は，主に日本語の自然言語で書かれたテキストである．そのテキストを解析し適切な形に変換する必要がある．テキストデータの可視化は主に次のよう手順で行う．(1) データ収集．学会の Web サイトに公開された論文題目等の情報を自動的に収集する．(2) テキストデータの前処理．準備したテキストデータをデータ解析に適したものに変換する．(3) テキストデータの可視化．前処理を施したテキストデータを可視化ツールで可視化する．(3) 可視化結果の確認・評価．可視化を実行し，年度，ジャンル別の変化を確認する．必要に応じて使用するデータの条件の見直し，前処理からやり直す．

2.1 テキストデータの前処理

コンピュータが取り扱うすべての情報は 2 進数の 0 と 1 のみであるため，テキストデータの多くはそのままの状態ではデータ解析を行うことはできない．また，テキストデータには句読点などの記号やデータ解析には不要なワードが多く含まれる．そのため，データ解析の精度を高めるためにもテキストデータを加工しなければならない．この処理を「前処理」という．本研究で行う前処理の流れは以下の通りである．

(1) テキストデータを読み込み，適切なデータ構造に格納する．
(2) 条件を指定しデータ構造から必要なテキストデータを抽出する．
(3) 抽出したテキストデータに対して形態素解析を行う．
(4) 形態素解析の結果から，特定の品詞やストップワードを除去する．
(5) 上記の結果を次の解析に適した形に変換する．

2.2 専門用語抽出

専門用語 とは，ある特定の職業に従事する者や，ある特定の学問の分野，業界等の間でのみ使用され，通用する言葉・用語群である．「テクニカルターム」とも言われる．学会等の学術的グループでの専門用語は特に「学術用語」と呼ばれる．英語では，Terminology, Technical Terminology, Technical Term 等の表現がある．

専門用語と一般語の区別は明確ではなく，各分野の専門家であれば明らかに専門用語であると判断できるが，新しい技術に関する用語が出てきた場合，共通して専門性を認識するまでには時間を要する．また，ある分野の初心者によってその単語が専門用語だとわかったとしても，どの程度の専門性であるのか，その分野における基礎的・必須用語なのか，分からない場合が多い [12][13].

専門用語は，(1) 言語学的な構造 (Linguistical Structure)，(2) 単位性 (Unithood), (3) 用語性 (Termhood) により決められている [3][5]．さらに，専門用語に，専門性 (Specificity, Technicality)，基礎性 (Fundamentality)，先端性 (State of the Art) という特徴がある [5][6][7].

専門用語は，単名詞だけではなく複合名詞になることが多い．複合名詞の例として，Judkins カテーテル，3 次元造影，人工多能性幹細胞などがあげられる．テキストから，複合名詞を識別するため，複合名詞の文法的構造，特に品詞のつながり方を示す必要がある．

日本語解析システムで使われている品詞体系は，IPADic 品詞体系と UniDic 品詞体系がある．IPADic 品詞体系の場合は，複合名詞の一部の品詞は，名詞，接頭詞，形容詞を基本とする．名詞のうち，さらに一般，固有名詞，サ変接続，ナイ形容詞幹，副詞委可能，形容動詞語幹，数，接尾等の中分類が含まれる．形容詞は自立のみ，接頭詞には，名詞接続，数接続が含まれる．

一方，UniDic 品詞体系の場合は，複合名詞を構成する部品の品詞は，名詞，接尾辞，形状詞を基本とする．名詞に

は，普通名詞, 固有名詞, 数詞が含まれる．接尾辞には，”名詞的”接尾辞,”形状詞的”接尾辞，形状詞には”一般”形容詞が有効である．

2.3 キーワード可視化

キーワード可視化とは，テキストマイニングの一種である．テキストマイニング（Text Minig）は，文字列を対象としたデータマイニングのことである．通常の文章からなるデータを単語や文節で区切り，それらの出現の頻度や共出現の相関，出現傾向，時系列などを解析することで有用な情報を取り出す，テキストデータの分析方法である．

テキストデータの多くは形式が定まっておらず，また日本語は英語などと比べて単語の境界判別の必要性（分かち書き）や文法のゆらぎが大きい点において形態素解析が困難であったが，自然言語処理の発展により実用的な水準の分析が可能となった．テキストマイニングの対象としては，顧客からのアンケートの回答やコールセンターに寄せられる質問や意見，電子掲示板やメーリングリストに蓄積されたテキストデータなどがある．

キーワード可視化の手法について，いくつか下記で説明する．

2.3.1 ワードクラウド

ワードクラウド（WordCloud）は文章中で出現頻度が高い単語を複数選び出し，その頻度に応じた大きさで図示する手法．ウェブページやブログなどに頻出する単語を自動的に並べることなどを指す．文字の大きさだけでなく，色，字体，向きに変化をつけることで，文章の内容をひと目で印象づけることができる．

2.3.2 棒グラフ

棒グラフとは，縦もしくは横軸にデータ量をとり，棒の長さでデータの大小を表したグラフである．値の高い項目や低い項目を判別するのに有効なグラフで，データの大小が棒の高低で表されるため，データの大小を比較するのに適している．キーワードの可視化のみならず，多くの場面で使用されている．

2.3.3 ThemeRiver

ThemeRiver とは，要素の時間的推移を川の流れのように提示する可視化手法で，横軸で時間を表現し，各要素を色で，各要素の値の大きさを垂直方向の幅で， 複数の要素の時系列変化を積み重ねて表示する [9][10][2]．この手法は，値の大きさが塗り分けの幅に対応しているため，どの要素が大きく変化しているかをユーザは一目で知ることができる．ThemeRiver は積み上げグラフの 1 種であり，経時的変化を表示する特殊なフロー グラフで横軸に沿って対称な可視化が行われる．

各トピックは「流れ」のように扱われ，離散的な時点の間を「流れる」．このようにして，各トピックはグラフ全体を通して一つの実体として整合性を保つ．離散的な時点から連続性を得るために，データポイントは曲がりくねった川のような柔らかい曲線に補間される．ThemeRiver は，大量の文書集合の中から傾向やパターンを特定し，トピックやトピックの予期せぬ発生や消失を発見することを可能にする．

3. データ収集

Web スクレイピングを使って，日本生体医工学会論文検索サイト[*1]より，2005 から 2023 年の間に掲載された学術論文 5,755 件を著者リスト，題目，出版年，巻，号を収集した．Web スクレイピングとは，インターネット上の Web ページから必要な情報を抽出・収集する技術である．プログラムが人間の代わりに Web ページを”閲覧”” しながら，対象となるコンテンツを自動的に取り出して保存する．Selenium はブラウザー操作の自動化を可能にする Python ライブラリーである[*2]．Selenium は以下の 3 つの機能が備えている．

(1) **ウェブページの操作**: ブラウザを開いてウェブページを表示したり，リンクをクリックしたり，テキストを入力したりすることができる
(2) **ページ内の要素の取得**: ウェブページの特定の要素（例えば，テキストや画像）を取得して，それを利用できる
(3) **動的なウェブページの操作**: JavaScript や Ajax などを使った動的なウェブページも，操作できる

Selenuim は，Webdriver を通して Web ブラウザーを操作することで Web 検索やクリック，情報の抽出や画面キャプチャを撮って保存することなどができる．今回は，Webdriver が Chrome ブラウザーより日本生体医工学会論文検索サイトの検索画面を操作することで検索結果を以下のように解析し，一つずつ論文題目，著者リスト，発表年，巻号，ページ数等を収集する．解析に利用する HTML タグは，次のようになっている．

- 検索結果全体を囲むタグ
  `<div id="search-resultslist-wrap">..</div>`
- 検索結果を箇条書きにするタグ
  `<ul class="search-resultslisting">`
  `<li>..</li></ul>`
- 論文題目を示すタグ
  `<div class="searchlist-title">..</div>`
- 著者リストを示すタグ
  `<div class="searchlist-authortags">..</div>`
- 出版年・巻号・ページ数
  `<div class="searchlist-additional-info">..</div>`
- DOI リンク
  `<div class="result-doi-wrape">..</div>`
- ダウンロードリンク

[*1] https://www.jstage.jst.go.jp/browse/jsmbe/list/-char/ja
[*2] https://selenium-python.readthedocs.io/

表 1: 生体医工学会論文の年度別論文件数

年度	論文誌巻号	論文数	年会	発表件数	題目総数
2005	43	96			96
2006	44	91			91
2007	45	36			36
2008	46	101			101
2009	47	75			75
2010	48	70			70
2011	49	144			144
2012	50	73			73
2013	51	60			60
2014	52	454			454
2015	53	678			678
2016	54	9	Annual54	454	563
2017	55	41	Annual55	365	406
2018	56	36	Annual56	423	459
2019	57	39	Annual57	568	607
2020	58	38	Annual58	407	445
2021	59	32	Annual59	551	583
2022	60	34	Annual60	356	390
2023	61	10	Annual61	483	493
不明	不明	2			2

```
<div class="global-tags">..</div>
```

最終的に表 1 に示すように 2005 年度から 2023 年度まで計 19 年間，論文誌，年会発表を含む題目が計 5,755 件収集できた．表を見てわかるように，2005〜2013 の間には，年間論文件数が 100 件前後に対し，2014 以降は，論文件数が急速に増加したことがわかる．また，2014〜2015 の 2 年間は論文誌と年会を区別せずに登録されているが，2016 以降は Annual54 のように年会が学会誌と別に登録されているようにみえる．

表 2 に収集されたデータの例を示している．今回の解析対象は論文題目だが，巻号，著者リスト，年度，ページ数等の情報も収集された．論文題目は日本語だけではなく，英語，記号等も含まれているが，解析は主に日本語を対象とした．

4. 専門用語抽出

4.1 接続情報に基づく用語抽出

専門用語は複合語の形になることが多く，テキストから複合語を適切に抽出することが専門用語抽出に必要である．複合語の重要度スコアは単名詞の左右にほかの単語と接続する．中川らはこの接続情報を利用して複合語を抽出する方法を提案した [14][15]．まず，単名詞 N の左方接続頻度 LN と右方接続頻度 RN を求める．そして，単名詞の左右の接続頻度に基づき，複合語のスコアを計算する．単名詞 $N_1, N_2, \cdots, N_L$ が この順で連接した複合名詞を CN とする．CN のスコアとして各単名詞の左右のスコアの平均（ここでは相乗平均を採用する）を取り，CN の長 さに

表 2: 生体医工学会論文題目例

巻	号	論文題目
48	1	α 波を生理指標とした覚醒度と身体動揺との関係
48	1	UVB 照射によるマウス皮膚微小血管床における急性炎症反応に関する研究
48	1	ポータブル超音波診断装置に搭載可能な心臓の 3 次元計測システムの開発
49	1	複数の機械学習手法を用いた退院時サマリからの自動 DPC コーディング
49	1	腫瘍識別器の Leave-One-Out による性能評価結果の信頼性に関する考察
54	1	ヒト肺亜細葉 4 D モデルによる肺拡散能シミュレーション
54	1	近赤外レーザードップラー流速分布計測装置の皮膚癌診断と治療への応用
54	1	センサ付把持鉗子の把持面間に存在する腫瘍サイズの推定法
56	1	電気インピーダンスの周波数特性計測による細胞の生存判別
56	1	対極板の違いにより発生する低周波雑音が術中モニタリングに与える影響

依存しないスコア $LR(CN)$ を式 (1) のように定義できる．

$$LR(CN) = (\prod_{i=1}^{L}(LN(N_i)+1)(RN(N_i)+1))^{\frac{1}{2L}} \quad (1)$$

さらに，tf × idf スコアと同様に，用語候補である単名詞あるいは複合名詞が単独で出現した頻度を考慮すべく，式 1 を補正し，次のように $FLR(CN)$ を定義する．

$$FLR(CN) = f(CN) \times LR(CN) \quad (2)$$

また，「異なり数」，「パープレキシティ」のような単名詞の連接情報は，連接した単語の種類をカウントする方法（異なり数），（パープレキシティ）の方法がある．単名詞 N の左方接続種類数（「異なり数」）LDN と右方接続種類 RDN とするとき，パープレキシティスコア PP は次のように定義する．

$$PP(CN) = (\prod_{i=1}^{L}(LDN(N_i)+1)(RDN(N_i)+1))^{\frac{1}{2L}} \quad (3)$$

「情報科学技術」における複合語抽出を考える．この語は 3 つの単名詞「情報」「科学」「技術」に分割できる．この際，表 3 に示すように，それぞれの単名詞が他の単名詞とどれだけ結びつくか統計的に分かっているとする．

式 1 によれば，CN =「情報科学」のとき，$LR(CN) = \{(1+1)(2+1)(2+1)(3+1)\}^{1/4} = 2.913$, CN =「科学技術」のとき，$LR(CN) = \{(2+1)(3+1)(1+1)(1+1)\}^{1/4} = 2.632$, CN =「情報科学技術」のとき，$LR(CN) = \{(1+1)(2+1)(2+1)(3+1)(1+1)(1+1)\}^{1/6} = 2.569$ となる．

表 3: 複合語重要度の計算例

単名詞 N	左方接続頻度 $LN(N)$	右方接続頻度 $RN(N)$
情報	1	2
科学	2	3
技術	1	1

4.2 共起関係グラフに基づくキーフレーズ抽出

キーフレーズ抽出は，文章からその主題を良く表現している句を抽出する技術であり，統計ベース，グラフベース，機械学習ベースの手法に大別され，キーワード抽出ともいえる．キーフレーズは「人工—呼吸—器」のように複数単語の連続を抽出するので，単語を意味する「ワード」ではなく，句を意味する「フレーズ」が使われる． グラフベースのキーフレーズ抽出手法が PageRank を根幹にして単語の共起関係ネットワークから，重要度を算出し，重要なフレーズを抽出している．本研究では，グラフベースのキーフレーズ抽出手法として MultipartieRank アルゴリズムを利用する．

MultipartieRank は TopicRank を改良したキーフレーズ抽出手法である．TopicRank の課題であるトピックレベルで順位付けされるため，同一トピックに属するフレーズ候補の集合は同じ重要度となり，代表フレーズがヒューリスティックに抽出されるため，完全な自動抽出ではない． MultipartieRank ではグラフ形成の仕方を工夫することで，この課題を回避している．今回はキーフレーズ抽出手法を実装したライブラリとして pke[*3]を使うことにした．pke は英語・フランス語等の欧米の言語の対応が基本で分かち書きや品詞推定には spaCy が用いられている．日本語対応のため，GiNZA ライブラリを利用する．表 4 は，MultipartieRank によって抽出されたキーフレーズの例である．

専門用語抽出は，これまで紹介してきた技術である程度実現可能だが，完璧に行うためには専門分野の深い知識と理解が必要不可欠である．表 6 は専門家により作成された生体医工学ウェブ辞書にある用語例である．形態素解析後，語と語の間に「.」を入れて語の構造は「説明」欄に説明している．

5. 可視化

5.1 ワードクラウドによるキーワード可視化

本研究でキーワード可視化を行うにあたり使用したライブラリは nlplot であり，nlplot の実行に必要なライブラリとして Plotly と Matplotlib がある．nlplot には頻出上位と頻出下位の指定できるストップワード機能があるが今回は使用していない．nlplot はいくつかの可視化の方法が用意されているが，本研究では Wordcloud を採用した．

nlplot は自然言語の基本的な可視化を手軽に行えるライブラリであり，日本語と英語で動作する[*4]．基本的な描写は Plotly を用いているため，ノートブック上からインタラクティブにグラフを操作することができる．nlplot で使用するデータの形式はデータフレームを想定しており，日本語を可視化する場合は事前に分かち書きをする必要がある．

表 4: spacy+pke によるキーフレーズ抽出例

キーフレーズ	重要度スコア
ヒト iPS 細胞 由来 心筋 細胞	0.003591
細胞 外電	0.003574
最適 設計	0.003573
ウェアラブル 心電図 計測 技術	0.003531
容量 結合 方式	0.003498
多元 計算 解剖 モデル	0.003474
空間的 配置 条件	0.003470
走査 方法	0.003444
超音波 音源	0.003424
内視鏡 シミュレーター ロボット	0.001887
心房 細動	0.001884
カラー オフセット csk	0.001882
VR 認知症 体験	0.001880
血液 浄化 療法	0.001870
機能的 MRI データ	0.001865
臓器 モデル	0.001858
血管 内皮 細胞	0.001856

表 5: Termextract による専門用語抽出例

専門用語	重要度スコア	出現頻度
fNIRS	7.416	14
モデリング	5.099	5
タンパク質	4.899	4
ヘルスケア	4.472	6
数理モデル	4.000	16
マイクロ波	3.742	3
自律神経系	3.162	5
プログラム	3.000	10
周波数特性	2.449	3
有限要素法	2.000	3
生体インピーダンス法	1.000	3
定常状態視覚誘発電位	1.000	4
ニューラルネットワーク	3.742	6
心臓リハビリテーション	1.000	3
コラーゲンゲルチューブ	1.000	3
レギュラトリーサイエンス	3.464	4
ダブルルーメンカテーテル	1.414	4

ワードクラウドによる可視化の結果と考察

時代の移り変わりに伴い，テーマの変化を確認したいので，2008 年度〜2015 年度の 8 年間を一つのグラフ，約 8 年後の 2016 年度〜2023 年度を別のグラフにして比較しやすいように左右に並べる．

[*3] https://github.com/boudinfl/pke

[*4] https://pypi.org/project/nlplot/

(a) 2008〜2009 年度

(b) 2010〜2011 年度

(c) 2012〜2013 度

(d) 2014〜2015 年度

図 1: 2006〜2015 年度頻出キーワードの WordCloud

表 6: 生体医工学ウェブ辞書例 (分ち書き後)

番号	用語名称	説明
001	BRS	英語・固有名詞
001	Bioresorbable vascular scaffold	英語・固有名詞
001	生体. 吸収. 性. スキャフォールド	複合名詞
002	Judkins. カテーテル	複合名詞
003	f-TUL	英語・固有名詞
003	r-TUL	英語・固有名詞
003	経. 尿道. 的. 尿管. 砕石. 術	複合名詞
004	iPS 細胞	固有名詞
004	人工多能性幹細胞	固有名詞
005	アナログ. 信号	複合名詞
005	アナログ信号処理	固有名詞
006	アナログ. 電気回路	複合名詞
027	カテーテルアブレーション	固有名詞
028	活動電位. 伝播	複合名詞
028	興奮. 伝播	複合名詞
029	株. 化. 細胞	複合名詞
030	冠動脈. 形成. 術	複合名詞
031	冠動脈. 造影	複合名詞
032	間葉系幹細胞	固有名詞

2008 年度〜2015 年度のワードクラウドは図 1 に示している．8 年間にわたって，「工学」といった大きなキーワードが継続しつつ，ほかのキーワードが大きく変わったことがわかった．特に「生体」，「信号」，「応用」，「膜」，「力学」等のキーワードが，「センサ」，「画像」，「NIRS」等のキーワードにとって代わり，また，「デバイス」や「波」といったキーワードがＡＩ人気に伴い，新たに表れた．

2016 年度〜2023 年度のワードクラウドは図 2 に示している．「センサ」以外，大きなキーワードがほぼすべて変わった．特に「光」，「ヒト」，「脈」，「心筋」等のキーワードが，「医」，「可視化」，「音」等のキーワードにとって代わり，また，「CT」や「抽出」といったキーワードが新たに表れた．

5.2 ThemeRiver による研究動向変化の可視化

時代の変化に伴い，新しい研究課題に直面し，新しい技術が開発されるため，それがどのように研究のトピックに反映されたかを調べることによって，研究動向の変化を把握できるかもしれない．

研究動向の可視化において大きなチャレンジとしては，可視化の対象となる**トピックの選定**と，トピック毎の**統計情報の獲得**が挙げられる．特に専門性の高い内容こそ，専門用語のレベルが高くなれば，統計情報が少なくなり，可

(a) 2016〜2017 年度

(b) 2018〜2019 年度

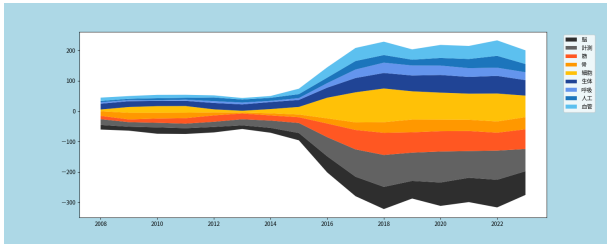
(c) 2020〜2021 年度

(d) 2022〜2023 年度

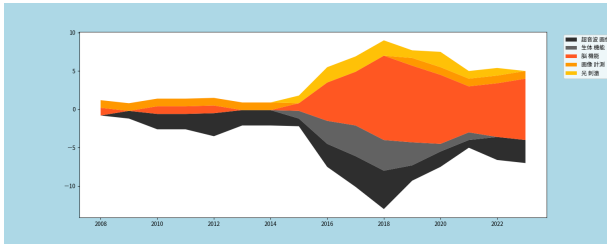
図 2: 2016〜2023 年度頻出キーワードの WordCloud

視化が難しくなる傾向がある．一般的に，論文題目として，用語の出現頻度が高いほど，専門性の低い一般語になりがちである．

本研究では，Termextract 法で抽出した用語のうち，代表的な単名詞（Uni-gram トピック）と長さ 2 の複合名詞 (Bi-gram トピック) を選定し，年度別出現頻度を調べる．その結果を ThemeRiver で可視化し，図 3(a) と (b) に示している．Uni-gram トピックは単名詞なので，様々な用語の一部になりえるため，出現頻度が数十と高い．しかし，Bi-gram トピックになると，ほかの用語の一部になるケースが大幅に減り，出現頻度が一桁になってしまう．



(a) Uni-gram トピックの ThemeRiver



(b) Bi-gram トピックの ThemeRiver

図 3: トピックの ThemeRiver

ThemeRiver は積み上げ面グラフとして描画した．積み上げ面グラフは基本的な面グラフを拡張したもので，各グループの値が重なって表示されるので，数値の合計と変化を同じグラフ上で確認することができる．

Python の Matplotlib で stackplot 関数を使って積み上げ面グラフを作成することができる．stackplot 関数は以下のように定義されている．

```
stackplot(x, *args, labels=(), colors=None,
    baseline='zero', data=None, **kwargs)
```

引数は次のようになる．

- x (配列)：横軸となる 1 次元配列．
- $args$ (可変長位置引数): 縦軸となるデータ（2 次元配列か 1 次元配列の繰り返し）．2 次元配列の場合は，行の数がグループの数，列の数が x の要素数と同じ．

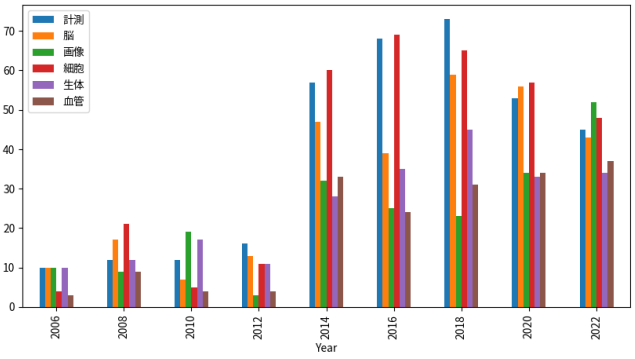


図 4: 代表トピックの出現頻度の推移

- *labels* (文字列の配列): ラベル文字列の配列．要素数が縦軸のデータ数と同じ．
- *baseline* (文字列): ”zero”, ”sym”, ”wiggle”などを指定することで面グラフの形式に変更可能 .
- *colors* (色の配列): 要素数が縦軸のデータ数と同じ．

stackplot 関数の引数を $baseline =' sym'$ とすると，ゼロを中心に対称な積み上げ面グラフを描画でき，ThemeRiver になる．また，$baseline =' wiggle'$ とすると，ストリームグラフが描画される．ストリームグラフは中心軸を軸として層の「揺れ」が最小になるように配置される．

ThemeRiver と比較するため，代表的な専門用語の年度別の出現度を棒グラフとして図 4 を描画した．同じ傾向が観察できるが，ThemeRiver のほうがより視覚的に分かりやすいのではないかと思われる．また，2013 年度まではデータ総数が少なかったため，出現頻度もそれに相応する結果になっている．

6. 終わりに

本研究では，生体医工学の 19 年間にわたる学術情報（計 5,755 件）を収集し，テキスト解析と可視化を試みた．この試みを通じて，研究テーマがどのように形成され，時代の変遷に伴ってどのように変化してきたのかを細部にわたって把握することが可能となった．こうしたアプローチから得られた知見は，データ可視化の手法や課題についての理解を一層深めるものとなった．

従来，「ツールを使えば，可視化はだれでも簡単にできる」とされていましたが，本研究を通してその限界も明らかになった．特に，専門用語抽出などの前処理作業において，ツールだけでは難易度が高い課題も浮かび上がった．これは，データの複雑性や多様性が高まるにつれ，専門的な処理や深い理解が求められることを示唆している．

今後の研究においては，現行の成果を基にしていくつかの重要な方針や取り組みを計画している．今後の課題として，特徴量抽出によるトピックの自動抽出ができたが，専門性が低い単語がまだ頻出上位に表示していることを改善していきたいと考えている．

同時に，先端技術や新たな手法の採用にも焦点を当てる．特に自然言語処理の最新動向や他の学問分野からの知見を取り入れ，データ解析手法の精度向上を目指したいと考えている．これにより，研究動向の可視化においてより包括的で洗練された手法を提供できると期待している．

参考文献

[1] J. Liu, T. Tang, W. Wang, B. Xu, X. Kong and F. Xia, *A Survey of Scholarly Data Visualization*, in IEEE Access, vol. 6, pp. 19205-19221, 2018, doi: 10.1109/ACCESS.2018.2815030.

[2] Wang, J., Li, Z. and Zhang, J. *Visualizing the knowledge structure and evolution of bioinformatics*, BMC Bioinformatics 23 (Suppl 8), 404 (2022). https://doi.org/10.1186/s12859-022-04948-9

[3] Frantzi K, Ananiadou S, and Mima H. *Automatic recognition of multiword terms: the c-value/nc-value method*, International Journal on Digital Libraries, 2000, 3(2):115–130

[4] Justeson J S, and Katz S M. *Technical terminology: some linguistic properties and an algorithm for identification in text*, Natural Language Engineering, 1995, 1(1): 9–27.

[5] Fumimaro Odakura et al, *Active Learning for Extracting Technical Terms Covering Multiword Phrases*, iiWAS2021, pp. 311-318, https://doi.org/10.1145/3487664.3487706

[6] Zoubin Ghahramani and Katherine A Heller. 2005. *Bayesian sets*. Advances in neural information processing systems 18 (2005), 435-442.

[7] Nisha Ingrid Simon and Vlado Keselj. *Automatic term extraction in technical domain using part-of-speech and common-word features*, In Proceedings of the ACM Symposium on Document Engineering 2018. 1-4.

[8] Anna Hatty, et al, *Predicting Degrees of Technicality in Automatic Terminology Extraction*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2883-2889, July 5 - 10, 2020

[9] Byron, Lee, and Martin Wattenberg. *Stacked graphs?geometry & aesthetics*, IEEE transactions on visualization and computer graphics 14.6 (2008): 1245-1252.

[10] S. Havre, B. Hetzler and L. Nowell, *ThemeRiver: visualizing theme changes over time*, IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings, Salt Lake City, UT, USA, 2000, pp. 115-123, doi: 10.1109/INFVIS.2000.885098.

[11] 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.

[12] 内山 清子, 専門用語の専門性判定に関する一考察，Japio YEAR BOOK, 2010. pp.152-153

[13] 内山 清子, 専門分野における用語の分野基礎性に関する研究, 言語処理学会 第 17 回年次大会 発表論文集 (2011 年 3 月)

[14] 中川 裕志 等, 日本語マニュアル文における名詞間の連接情報を用いたハイパーテキストのための索引の抽出,『情報処理学会論文　第３８巻　第１０号』平成 9 年 10 月.

[15] 中川 裕志 等, 出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理, 2003, 10 巻, 1 号, p. 27-45, https://doi.org/10.5715/jnlp.10.27