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

張　馨雲[1]　今泉 優気[1]　隈部 晶[1]　林　成元[1]　豊坂 祐樹[2]　成　凱[1,a)]

概要：近年，学術情報の電子化・大規模化が急速に進み，学術データに隠れたパターンを解析し，視覚的に確認することが研究者にとって重要になっている．本研究では，卒業研究のテーマを解析し，その中で隠れた技術分野の変遷を可視化し，学生，教員に提示することで，学生の研究室選択や教員の研究指導に役立つ事を目的とする．具体的には，情報科学科約 20 年にわたる卒業研究のテーマに対して形態素解析を行い，出現頻度の高い単語をキーワードとしてワードクラウドを使って可視化する．また，年度別，研究室別に集計を行い，それぞれのキーワードの変化を ThemeRiver で可視化する．実験結果として，教員の研究分野に沿った結果であったものの，それぞれの研究室で時代に伴いテーマの変遷が確認できた．

Visualizing Trends in Graduation Research Topics by Text Analysis and Machine Learning

1. はじめに

近年，学術情報の電子化・大規模化が急速に進み，学術データの分析を通じて重要な洞察を引き出すとともに，データを視覚的に提示すること，いわゆる「データ可視化」が重要になっている [3][4]．データ可視化（Data visualization）とは，数値情報だけでは確認しにくい現象や関係性，変化などを一目見れば分かる形（グラフ，チャート，表，画像等）に変換し，データに隠れた情報を表示して，数字から分かる情報の理解を助けることである．データの「見える化」や「視覚化」とも呼ばれる．可視化により，大規模なデータセットのパターン，傾向，関係性，外れ値を簡単に識別できる．

しかし，一言「可視化」といっても，分野，解析目的やデータセットに応じて，データ処理方法や可視化方式を適切に決めることは簡単ではない．まず，データには，数値だけではなく，日付，テキスト，カテゴリ等も含まれている．また，データがある静止状態を表すだけではなく，時間に伴う移り変わりや傾向を可視化することも重要である．さらに，大規模で複雑なデータを扱うとき，可視化の対象を適切に決めることが容易ではない．

本研究の目的は，情報科学科で暦年行われた卒業研究のテーマを対象とし，研究室間の研究内容や技術の変遷を可視化することで，学生の研究室選択や教員の研究指導に資するとともに，学術データの適切な可視化のノウハウを得ることである．

情報科学科には 16 ほどの研究室が存在し，３年次の学生は卒業研究のために研究室に配属され教員の指導を受けながら，４年次に１年間かけて卒業研究を行う．研究室配属では，学生に対して配属調査を３年次の前期に行い，その調査結果を基に配属先を決定している．その際に学生は各研究室が簡単に紹介され，２週間程度の期間で学びたいテーマがある研究室をいくつか決めて希望を提出している．

現在の研究室配属は一部の研究室に人気が集中してしまう傾向があり，その結果抽選という形となり自分が希望した研究室に配属されないケースが多く存在する．これは研究室を決める期間が２週間程度と短いため，自分に適した研究室を選択できているかどうかわからず，研究室を十分に理解せずに希望を提出していると考えられる．また，４年次になり卒業研究を開始しても学生が研究したいものとは異なるテーマになってしまい研究について行けず，除籍や退学になってしまうこともある．

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

© 1959 Information Processing Society of Japan

さらに，教員が学生を指導するに当たって適切な研究テーマを決めることが重要である．研究テーマを決めるために，教員自身の研究分野に関連するテーマを学生に繰り返しやらせるだけではなく，学問分野や関連技術の発展や他研究室指導の傾向を把握することも重要な参考になる．

本研究では，情報科学科の約 20 年にわたって行われた卒業研究のテーマを対象に，形態素解析を行い，出現頻度の高い単語をキーワードとしてワードクラウドを使って可視化する．また，年度別，研究室別に集計を行い，それぞれのキーワードの変化を ThemeRiver で可視化する．これにより研究室毎の研究テーマ，キーワードの違いを明確にし，３年次の研究室配属や４年次の卒業研究のテーマ選択において，学生と研究室のミスマッチを減らすことが期待できる．

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．本研究の対象となった卒業研究テーマは，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) 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．(2) 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) 年度や研究室等の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 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．

2.2.1 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．

2.2.2 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．

2.2.3 共起ネットワーク

共起ネットワークとは，テキストに含まれる単語間の共通点を見つけ，図で表現する手法である．テキストにおける単語同士のつながりを可視化し，視覚的に理解を促せるため，テキストマイニングの手法として非常に人気である．

2.2.4 サンバーストグラフ

サンバーストグラフ（Sunburst Chart）は，階層データの表示に最適で，階層の各レベルを 1 つのリングが表し，最も内側の円が階層の最上位に相当する．複数レベルのカテゴリを持つサンバーストグラフは，外側のリングと内側との関係を示す．

2.2.5 ツリーマップ

ツリーマップ（Tree Map）とは，二次元平面上の領域を入れ子状に分割することによって，木構造のデータを可視化する手法．2 つの大きな利点を持っている．

1 つ目は，空間効率の良い情報可視化を実現できる．画面を隙間なく利用するため，限られた空間に多くの情報を詰め込むことができる．2 つ目は，分割された各領域の面積を自由に決められる．

© 1959 Information Processing Society of Japan

2.2.6 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 [1][2][4]．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．

ThemeRiver は，川のメタファーを使って一定期間の文書集合におけるテーマ別の変化を視覚化したものである．横軸は時間の連続単位で，縦軸は横軸の時間において文テーマの ”強さ”を表している．ヒストグラム図に似ているものだが，ヒストグラムは，複数の棒グラフをデータとともに積み重ね，それぞれがタイムスライスを表す一般的な可視化手法である．それに対して ThemeRiver は，連続的な時間を川のメタファーで表現する．

各テーマは「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，各テーマ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，テーマ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. ワードクラウドによる可視化

本研究では，学内ページから確認できる平成 17 年度以降の情報科学科卒業生の卒業論文テーマに対して形態素解析を行い，出現頻度の高い単語をキーワードとして可視化する．また，年度別，研究室別に可視化を行い，それぞれのキーワードの変化を確認する．

実験では，出現頻度の高い単語をキーワードとして可視化する．このとき年度別，研究室別に可視化を行い，それぞれのキーワードの変化を確認する．すべての実験において，WordCloud に表示する語数は 50 となっている．まず，年度別のグループを比較し，研究テーマに利用されるキーワードの変化を調べる．1 年分のテキストデータでは量が不十分であることから，結果に偏りが出てしまう可能性があった．そのため複数年分を 1 つのグループとし，グループ別に可視化し，キーワードの変化を調べる．そして，研究室別のグループを比較し，研究テーマに利用されるキーワードの変化を調べる．ここでは平成 17 年度から令和 4 年度までのデータが存在し，かつ 100 以上の論文を出している研究室を対象に可視化し，キーワードの変化を調べる．同研究室内での年度別のグループを比較し，研究テーマに利用されるキーワードの変化を調べる．

表 1: 卒業研究テーマ例

年度	研究室	題目
H17	○○研	WEB ページにおけるユーザビリティの追及
H17	○◇研	安全運転管理教育システム (ASSIST) における複数カメラ使用での交通事故防止対策
H17	△△研	電子透かしの体制に関する検討 -加重平均フィルタ・メディアンフィルタの場合-
H17	◇○研	対面教育を支援するためのウェブ助言システム
H17	◇◇研	衛星通信検討のための気象データの解析-2002 年のデータ解析-

3.1 データの読み込み

本研究で使用するテキストデータは，理工学部情報科学科学生平成 17 年度以降に卒業した情報科学科の学生のものである．すべての学生の年度，学科／研究室，学籍番号，賞，氏名，要旨，題目を csv ファイルで保存して使用している．表 1 はサンプルを示している（スペースの関係上，一部の項目しか表示していない）．JupyterLab のデータが保存されているフォルダを保存先にすることで JupyterLab でのテキストデータの導入が簡単になる．データ量は平成 17 年度から令和 4 年度の間に卒業した情報科学科の学生の論文，計 2,148 件である．

3.2 テキストデータの前処理

本研究では一部の品詞のみを抽出し，それら以外の品詞のワードをストップワードに指定した後，特定のワードや記号を追加でストップワード処理している．また，使用する形態素解析エンジンは MeCab とし，辞書は新語・固有表現に強い「mecab-ipadic-NEologd」を使用した．実験対象となる品詞は動詞，名詞，形容詞である．

形態素解析例 1: 以下のテキストを解析とする．

「人感センサーを利用した乗降客数計測 *Android* アプリケーションの開発」

IPADIC 辞書を利用した解析結果は [’ 人’, ’ 感’, ’ センサー’, ’ 利用’, ’ する’, ’ 乗降’, ’ 客数’, ’ 計測’, ’Android’, ’ アプリケーション’, ’ 開発’] となり，「人感」という単語がうまく認識できなかった．一方，IPADIC-NEologd 辞書を利用した解析結果は [’ 人感センサー’, ’ 利用’, ’ する’, ’ 乗降客数’, ’ 計測’, ’Android’, ’ アプリケーション’, ’ 開発’] となり，うまく認識できた．

形態素解析例 2: 以下のテキストを解析とする．

「研究室配属における学生の研究室に対する理解を深める情報共有システム」

IPADIC 辞書を利用した解析結果は [’ 研究’, ’ 室’, ’ 配属’, ’ 学生’, ’ 研究’, ’ 室’, ’ 理解’, ’ 深める’, ’ 情報’, ’ 共有’, ’ システム’] となり，「研究室」や「情報共有」のような言葉が理解できなかった．一方，IPADIC-NEologd 辞書を利用した解析結果は [’ 研究室’, ’ 配属’, ’ 学生’, ’ 研究室’, ’ 理解’, ’ 深める’, ’ 情報共有’, ’ システム’] となり，うまく認識

© 1959 Information Processing Society of Japan

(a) 平成 19 年度～平成 22 年度　　(b) 令和元年度～令和 4 年度

図 1: 基礎系研究を行う研究室 A のワードクラウド

(a) 平成 19 年度～平成 22 年度　　(b) 令和元年度～令和 4 年度

図 2: ハードウェア関連研究を行う研究室 B のワードクラウド

できた．

形態素解析例 3: 以下のテキストを解析とする．

「拡張現実感のための実物体の奥行きを考慮した陰面消去」

IPADIC 辞書を利用した解析結果は [' 拡張', ' 現実', ' 感', ' ため', ' 物体', ' 奥行き', ' 考慮', ' する', ' 陰', ' 面', ' 消去'] となって，「拡張現実」という言葉が認識できなかった．一方，IPADIC-NEologd 辞書を利用した解析結果は [' 拡張現実', ' 感', ' ため', ' 物体', ' 奥行き', ' 考慮', ' した陰', ' 面', ' 消去'] となり，改善できたが，「した陰」のような誤判断もあった．

3.3 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[*1]．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．

3.4 ワードクラウドによる可視化の結果と考察

研究テーマは分野によってパターンが異なるのではないかと考え，主な分野を 4 つピックアップして紹介する．これらは，基礎系，ハードウェア系，ソフトウェア・ＡＩ系，ネットワーク系である．また，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，平成 19 年度～平成 22 年度の 4 年間を一つのグラフ，約 10 年後の令和元年度～令和 4 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．

基礎系の研究を行う研究室Ａのワードクラウドは図 1 に示している．10 年間にわたって，「シミュレーション」と

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

© 1959 Information Processing Society of Japan

(a) 平成 19 年度～平成 22 年度　　(b) 令和元年度～令和 4 年度

図 3: ソフトウェア・AI 関連研究を行う研究室 C のワードクラウド

(a) 平成 19 年度～平成 22 年度　　(b) 令和元年度～令和 4 年度

図 4: ネットワーク関連研究を行う研究室 D のワードクラウド

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「数学教育」,「ビーム」,「通信」,「衛星」,「Java」等のキーワードが,「マイクロ波」,「誘電体」,「Mathematica」等のキーワードにとって代わり，また,「ニューラルネットワーク」や「機械学習」といったキーワードがＡＩ人気に伴い，新たに表れた．

ハードウェア関連の研究を行う研究室Bのワードクラウドは図 2 に示している．「システム」以外，大きなキーワードがほぼすべて変わった．特に「KERNEL」,「タイマー」,「ロボット」,「制御」等のキーワードが、「HDL」,「FPGA」,「パワーコンディショナー」等のキーワードにとって代わり，また,「ARM」や「組込み」といったキーワードが新たに表れた．

ソフトウェア・AI 関連研究を行う研究室 C のワードクラウドは図 3 に示している．10 年間にわたって，ハードウェア系の研究室と同じく「システム」といった大きなキーワードほかにキーワードが大きく変わったことがわかった．「スケッチ」,「手書き」,「図形」,「アバター」,「OpenGL」等のキーワードが、「自動運転」,「車」,「画像」等のキーワードにとって代わり，また,「強化学習」や「DCGAN」といったキーワードがＡＩ人気に伴い，新たに表れた．

最後にネットワーク関連研究を行う研究室 D のワードクラウドは図 4 に示している．これまでと同じく 10 年間にわたり，研究テーマのキーワードが大きく変わった．「ネットワーク」,「コンピューティング」,「シミュレーション」,「プロトタイプ」等のキーワードが,「エッジ」,「コンテナライブマイグレーション」,「モバイルヘルスモニタリング」,「強化学習」等のキーワードにとって代わり，AI や IoT などの人気に伴った結果と推測される．

4. ThemeRiver によるトピック変化の可視化

4.1 トピックの抽出

トレンドを可視化するために，表示対象のトピックを膨大な数の候補から適切に選出することが重要である．一般的には，トピックがキーワードより大きな意味単位を使う必要がある．例えば，個別のキーワード「教材」,「開発」がトピックとすると，漠然なイメージがあり，技術的な傾向を示すためには,「教材　開発」のような大きな意味単位

© 1959 Information Processing Society of Japan

表 2: 2-gram トピック候補グループ 1

2-gram トピック候補	出現頻度
「安全運転　管理教育」	20
「管理教育　システム」	20
「システム　ASSIST」	20
「ドライビング　シミュレータ」	16
「管理　システム」	16
「ソフトウェア　開発」	14
「システム　構築」	13
「交通　標識」	11
「衛星　通信」	9
「問題　メタヒューリスティクス」	9
「メタヒューリスティクス　実験的」	9
「実験的　解析」	9
「システム開発　演習」	9
「データ　解析」	8
「システム　試作」	8
「組合せ最適化　問題」	8
「記録　システム」	8
「クラスタ　コンピューティング」	8
「電磁波　伝搬」	8
「機能　開発」	8

表 3: 2-gram トピック候補グループ 2

2-gram トピック候補	出現頻度
「標識　抽出」	7
「システム　設計」	6
「講義　記録」	6
「コンピューティング　教育支援」	6
「伝搬　シミュレータ」	6
「認識　研究」	6
「省エネルギー 対策」	5
「サーバ　構築」	5
「システム　作成」	5
「セマンティック　グリッド」	5
「運転　行動」	5
「時間　記録」	5
「教育支援　環境」	5
「声道　模型」	5
「マルチ　ビーム」	5
「挙動　解析」	5
「経路　探索」	5
「業務　プロセス」	5
「シミュレータ 開発」	5

を採用すべきである．

本研究では，前述の方法で抽出されたキーワードから，単語 n-gram を求め，出現頻度をカウントして，出現頻度の高いものからトピックを選定する．H17 年度から H20 年度まで全 16 研究室 4 年間の 2-gram をさらに調べ，出現頻度の高い順に出力した．表 2 と表 3 は，その結果を示している．

上記の結果を見ればわかるように，「教材 作成」，「安全運転 助言」，「標識 識別」，「Web アプリケーション」等がトピックとして適切なものがある一方，「メタヒューリスティクス 実験的」，「助言 検査」等，わかりにくい組み合わせも出てくるので，トピックとして適切ではないものを除去する必要がある．

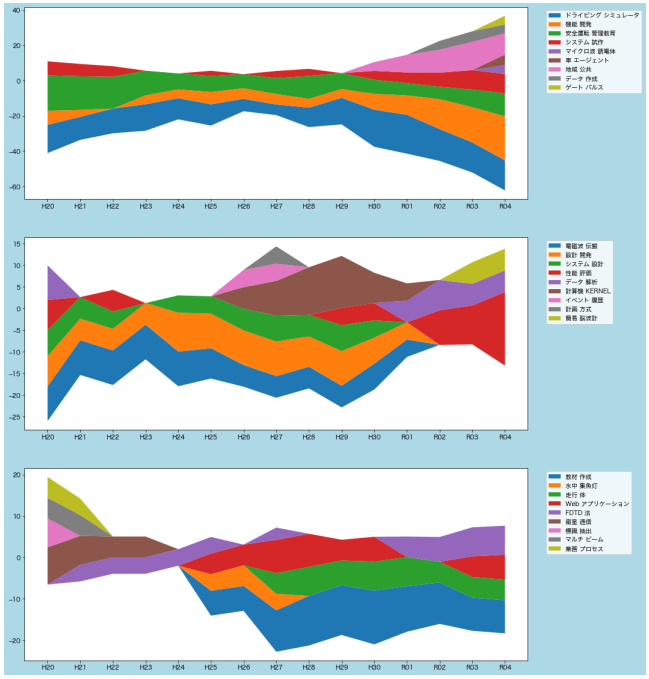

図 5: 代表的な 27 トピックの ThemeRiver

今回はトピックの存続する異なるパターンを考慮して，著者らの判断でトピックを選ぶことにした．選ばれたのは以下の 27 トピックである：「ドライビング シミュレータ」，「機能 開発」，「安全運転 管理教育」，「システム 試作」，「マイクロ波 誘電体」，「車 エージェント」，「地域 公共」，「データ 作成」，「ゲート パルス」，「電磁波 伝搬」，「設計 開発」，「システム 設計」，「性能 評価」，「データ 解析」，「計算機 KERNEL」，「イベント 履歴」，「計画 方式」，「簡易 脳波計」，「教材 作成」，「水中 集魚灯」，「走行 体」，「Web アプリケーション」，「FDTD 法」，「衛星 通信」，「標識 抽出」，「マルチ ビーム」，「業務 プロセス」．

4.2　stackplot による ThemeRiver の描画

ThemeRiver は積み上げ面グラフとして実現することができる．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．

© 1959 Information Processing Society of Japan

- 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.
- $labels$ (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.

4.3 ThemeRiver による可視化の結果と考察

ここまで述べてきた方法で，選定した 27 のトピックについて，図 5 に示すような ThemeRiver を描画した．横軸は年度を示しているが，直近 3 年間のテーマを合わせたものになる．例えば，令和 4 年度のデータは令和 2 からの 3 年間のデータを解析した結果になる．縦軸は，すべてのトピックの出現頻度を積み上げた状態にしたものでそれぞれのトピックの出現頻度は色わけの縦幅分で示している.

図 5 から以下の事実が読み取れる.

まず,「ドライビング シミュレータ」,「機能 開発」,「安全運転 管理教育」,「システム 試作」といったトピックが長年継続して研究が行われていたことが分かった.

また,「電磁波 伝搬」,「設計 開発」,「システム 設計」も長年継続で行われたが，直近数年になると，使わなくなった．逆に「教材 作成」,「走行 体」,「Web アプリケーション」も数年経った後に現れ，その後出現回数が多くなった.

ほか，途中数年しか現れなかった「水中 集魚灯」,「イベント 履歴」,「計画 方式」などのトピックもあるし，最近しかなかった「車 エージェント」,「地域 公共」,「データ 作成」,「ゲート パルス」などのトピックも存在する．さらに「データ 解析」,「性能 評価」などは，途中で一時消えていたが，また復活したトピックもある.

以上のことから，ThemeRiver による可視化が，時代の流れに伴うテーマの変化をわかりやすく表現できたと思われる.

5. 終わりに

本研究では，情報科学科 20 年間にわたって蓄積された卒業研究テーマ計 2,148 件を対象にテキスト解析と可視化を行い，研究室単位での研究テーマの特色や，時代の流れに伴う研究テーマの変化を確認できた．これによって，データ可視化の進め方や課題についてある程度理解を深めた．ツールを使えば，可視化はだれでも簡単にできる [5][6][7] といわれていたが，しかし，前処理の一環であるトピック抽出など，ツールだけでは簡単にできない部分もあることが研究を通してわかった．今後の課題として，n-gram の長さを指定せず，深層学習などを使ってトピックを抽出することなどがあげられる.

謝辞 本研究の遂行にあたり，卒業研究テーマの実データを使用した．データを提供してくださった学位論文等検索ホームページの制作者に感謝する．なお、本研究は，学生個人を特定できるものではなく研究室名にも匿名化処理を施した.

参考文献

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

[2] 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.

[3] 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.

[4] 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

[5] 三末 和男，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)

[6] 小久保 奈都弥，データ分析者のための Python データビジュアライゼーション入門，翔泳社 (2020/8/6)

[7] @driller, 小川 英幸 等，Python インタラクティブ・データビジュアライゼーション入門，朝倉書店 (2020/12/5)

© 1959 Information Processing Society of Japan