統計数理(2020)
第 68 巻 第 2 号 233–246

特集「Institutional Research と統計科学」
［原著論文］

大規模大学における研究分野の研究実績の可視化

船山 貴光[1]・山本 義郎[2]・藤野 友和[3]

（受付 2019 年 7 月 15 日；改訂 2020 年 3 月 10 日；採択 3 月 30 日）

要　　旨

大規模大学では，多くの研究者が在籍し，大小様々な規模で研究活動が行われている．また，研究領域は多岐に渡るため，大学全体の研究活動状況を把握することは困難である．学内研究者により活発に成果を挙げている研究領域を把握することは，評価の観点だけでなく学内研究助成や研究組織の構成などにも必要なことである．そこで本研究では，学術文献データベースに収録されている論文のタイトルとアブストラクトのテキストデータに対して，トピックモデルを用いてその論文の研究領域を推定し，研究業績の多い研究領域の把握を試みた．更にトピックモデルの結果を自己組織化マップを用いた可視化を行うことで，トピックモデルで分類された研究領域の特徴や研究領域間の関連性の把握ができることを示した．そして，自己組織化マップの結果を利用したいくつかの可視化を提案し，学内の研究傾向やその経時的変化を把握するための方法を例に示した．

キーワード：トピックモデル，自己組織化マップ，可視化．

1. はじめに

大学内の研究活動を支援及び評価をする上で研究活動状況の把握が必要不可欠である．しかし大規模大学の場合，研究者ひとりひとりの研究状況を把握することは研究分野により論文や著者の価値が異なるため困難である．学内でどんな分野が活発に研究業績を挙げているのかについて，様々な研究領域の研究が行われているため学部学科等の研究者が所属する組織単位での研究論文数を比較することで把握するのは困難であり，掲載されたジャーナルから分野を把握するのも困難である．そのため，学内研究者がどんな研究分野で業績を挙げているのかを把握するためには，蓄積された研究業績データから研究領域を推定する必要がある．本研究では，学術論文データベースを活用して，学内研究者が著者(共著も含む)である論文に関する情報から学内の研究実績の把握を試みた．

論文の研究領域について考える．一般的には，ジャーナルのタイトルなどになっている研究領域がその論文の研究領域とされる．しかし，複数の研究領域をカバーする研究成果の場合，カバーする研究領域を全て満たすことができないことがある．例えば，本稿のようにInstitutional Research(IR)における統計解析に関する研究が統計科学分野のジャーナルに投稿された場合，IR に関する研究であることはジャーナルのキーワードからは得られない．また，研究者の所属から研究領域を決定する方法も考えられるがこれも同様の問題が発生する．例え

[1] 東北大学 東北メディカル・メガバンク機構：〒 980–8573 宮城県仙台市青葉区星陵町 2–1
[2] 東海大学 理学部：〒 259–1292 神奈川県平塚市北金目 4–1–1
[3] 福岡女子大学 国際文理学部：〒 813–8529 福岡市東区香住ヶ丘 1–1–1

ば，本稿の場合である．本稿の著者は，それぞれ医学，理学，国際文理学の分野の所属であり，本稿の内容である統計学はそれぞれの所属分野に関係する研究領域であるが，IR はどの所属の研究領域からも連想することは難しい．このような理由から著者の所属学部やジャーナルの専門分野からその論文が扱っている研究領域を分類することは難しく，論文の内容からその論文の研究領域を決定することが必要となると考えられる．そのため本研究では，学術論文データベースに収録されている論文のタイトルとアブストラクトのテキストデータから論文の研究領域を推定した．研究領域の推定にトピックモデルを用いることで，各論文について各トピックのトピックへの所属確率を得られることから，それらの値に自己組織化マップを適用し，トピックの関連性について考察する．自己組織化マップを用いることで様々な可視化が可能となることを実例により示す．

本論文では，ケーススタディとして大規模な総合大学である T 大学を例に分析を行う．また，学術文献データベースは，クラリベイト・アナリティクス社の提供する Web of Science（WoS）を使用した．さらに，図 4 から図 8 は文字が小さくなってしまうため supplementary material を用意した．

2. データセットの作成

本研究では，提供されたデータを Neo4j によるグラフデータベースに格納し，データ管理及びデータの抽出を行なった．Web の WoS には，「著者の所属（大学名）」の変数があるが，提供された Neo4j データベースには無く，著者の所属を示す変数は，「著者の住所」のみであった．著者の住所とは，研究機関名（大学名など），部署（学部学科など），住所の順に記述された文字列データである．本研究では，著者の住所が “T 大学” から始まる著者を T 大学に所属する研究者とし，2007 年から 2016 年に発表された T 大学に所属する研究者（大学院生を含む）を著者に含む論文（4,261 編，以下では「T 大学著者論文」）とそれらの論文を引用した論文（36,604 編，以下では「T 大学引用論文」）のタイトルとアブストラクトを抽出し，T 大学著者論文と T 大学引用論文を結合したデータセットを作成した．データセットの論文のタイトルとアブストラクトのテキストデータを形態素解析し，過去形などの単語を原形に統一し，さらにピリオドやカンマなどの記号やストップワードは除外した．今回は，英語で書かれた論文データであるため，形態素解析ソフトウェアとしては，TreeTagger（Schmid, 1994, 1995）を使用した．以上の前処理後のデータセットは，総論文数は 40,865 で，データセット内で使われている単語数は 133,430 であった．T 大学著者論文の T 大学所属の著者は述べ 23,001 人（同一著者を含む）で住所から判別できた所属は，学部 20,390，センター・研究所 809，大学院 762，付属病院 625，短期大学 14，不明 401 であった．不明は，大学名以外に住所しか記述されていないものである．

3. トピックモデルを用いた論文の研究領域の推定

論文の研究領域を推定するため，トピックモデルを用いた．トピックモデルは，文章を分類する手法の 1 つであり，1 つの文章は，複数のトピック（話題）を持つと仮定し，それぞれのトピックに属する確率をモデル化した手法である．桂井 他（2015）は，日本の論文データベースである CiNii に登録されている論文に対して，LDA（Latent Dirichlet Allocation）（Blei et al., 2003）によるトピックモデルを用いて著者同定を行った．藤野 他（2016）は，分析対象の研究組織に所属する研究者名と同名の研究者が WoS において所属が示されていない論文について，学内データによって算出された著者の特徴ベクトルを用いて組織の研究者か否かについて判別を行なった．トピックモデルにおいては，様々なモデルが提案されているが本研究では，LDA を採用し，R の topicmodels パッケージ（Grün and Hornik, 2011）を用いた．LDA は，事前分布

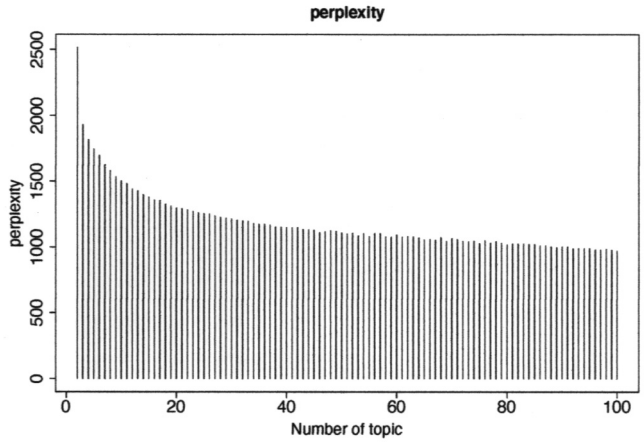

図 1．Perplexity の推移．

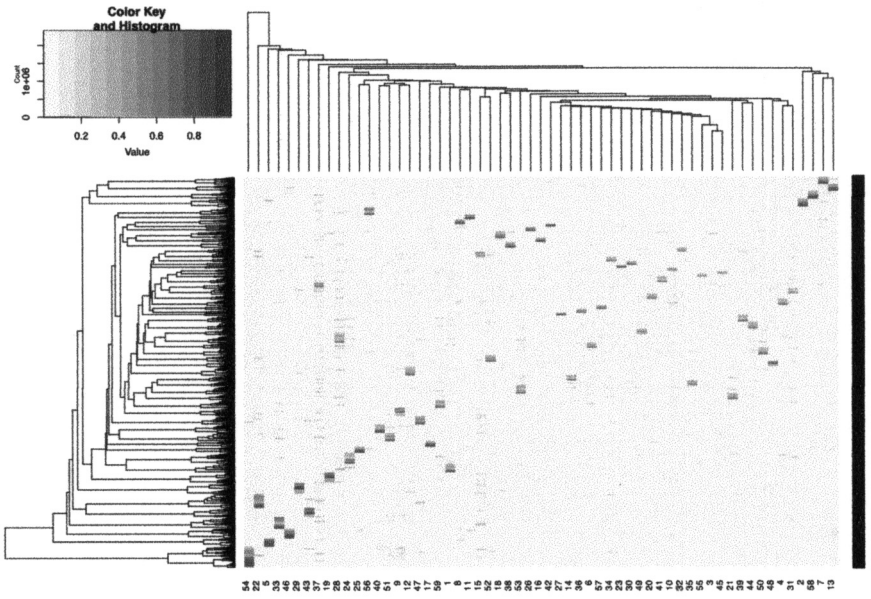

図 2．各論文のトピックの出現確率のヒートマップ．

に Dirichlet 分布を用いるため，特定のトピックへの所属確率が高くなる傾向がある．その性質を利用して最も所属確率が高いトピックに論文を割り当てることで各論文が 1 つの研究領域にふり分けられるようにした．トピックモデルは，事前にトピック数を設定する必要がある．本研究では，トピック数の評価指標の 1 つである Perplexity（Blei et al., 2003）を使用して妥当なトピックス数について判断した．Perplexity は，トピック数を増やすほど良い評価（小さな値）となる傾向があるが，トピック数が多過ぎると研究領域を極端に細分化してしまう．また，膨大な計算量になってしまうため実用的でなくなる．Perplexity が減少から増加に転じたトピック数，図 1 から 53，55，59，67，75 などを妥当なトピック数の候補と考えた．本論文では，計算コストを考慮してトピック数を 59 とした場合について論じる．LDA により推定された各トピックが示す研究領域は，各単語の出現確率が高い単語から判断することができる．図 2 は，

表 1. 各トピックの出現確率が高い上位 5 単語.

トピック番号	T 大学著者論文数	複数学部著者論文	研究領域	単語 1	単語 2	単語 3	単語 4	単語 5
トピック 1	43	19%	細胞	protein	mitochondrial	autophagy	mutant	function
トピック 2	55	4%	光工学	energy	telescope	gammaray	source	emission
トピック 3	3	0%	引用文献について	elsevier	right	reserve	reserved	publish
トピック 4	54	0%	乳癌・卵巣癌	cancer	tumor	breast	alpha	cell
トピック 5	59	5%	ニュートリノ	neutrino	matter	mass	dark	model
トピック 6	36	6%	細胞・遺伝子・ニューロン	cell	notch	development	signale	neural
トピック 7	162	0%	情報系・システム系	system	propose	proposed	use	method
トピック 8	103	1%	モデリング	model	use	function	result	method
トピック 9	83	0%	健康	age	patient	aging	study	care
トピック 10	38	11%	DNA・遺伝子	expression	gene	protein	express	analysis
トピック 11	90	1%	溶液を使っての実験	water	temperature	relaxation	dielectric	solution
トピック 12	62	0%	臨床医療	disease	patient	syndrome	disorder	clinical
トピック 13	184	1%	プラズマ・イオン・磁気	film	plasma	thin	ion	temperature
トピック 14	51	2%	骨・整形外科	bone	tissue	use	cartilage	fracture
トピック 15	71	0%	調査デザイン・調査の結論	card	patient	study	group	analysis
トピック 16	49	0%	遺伝・地理	population	asian	genetic	asia	human
トピック 17	78	0%	海洋	sea	water	ocean	high	surface
トピック 18	70	0%	ゲノム (基礎)	sequence	gene	class	allele	hla
トピック 19	100	0%	ゲノム (応用)	association	gene	study	polymorphism	genetic
トピック 20	44	7%	DNA・遺伝・マウス実験・ゲノム	dna	imprint	epigenetic	gene	methylation
トピック 21	80	0%	細胞・血管	cell	endothelial	vascular	epcs	progenitor
トピック 22	110	0%	心蔵血管	patient	risk	disease	stroke	cardiovascular
トピック 23	42	2%	地震	monitor	monitoring	earthquake	test	time
トピック 24	24	4%	オートファジー・細胞	autophagy	cell	induce	apoptosis	increase
トピック 25	98	5%	化合物・合成	center	dot	reaction	compound	synthesis
トピック 26	62	2%	気象・衛星	cloud	use	data	aerosol	satellite
トピック 27	19	0%	星	star	abundance	card	line	expose
トピック 28	11	0%	メカニズム	role	mechanism	disease	function	cell
トピック 29	165	1%	移植	patient	transplantation	cell	donor	leukemia
トピック 30	52	4%	筋・骨格・運動	muscle	skeletal	stimulation	motor	exercise
トピック 31	24	0%	薬剤治療	treatment	drug	therapy	target	therapeutic
トピック 32	108	0%	血液	group	blood	level	effect	significantly
トピック 33	190	1%	癌	cancer	patient	lung	treatment	survival
トピック 34	53	4%	脳・糖尿病・虚血	brain	rat	diabetes	injury	cerebral
トピック 35	47	0%	細胞	disc	degeneration	intervertebral	cell	ivd
トピック 36	94	1%	工場・化合物・毒	plant	compound	acid	extract	isolate
トピック 37	32	0%	関連研究についての説明	review	use	new	research	clinical
トピック 38	42	19%	ゲノム・遺伝子・進化	gene	vertebrate	evolution	genome	species
トピック 39	49	2%	腎臓	renal	kidney	nephropathy	injury	glomerular
トピック 40	109	2%	膵臓	case	pancreatic	tumor	lesion	neoplasm
トピック 41	53	0%	感染病	infection	virus	human	viral	marmoset
トピック 42	55	0%	自然・火山	delta	lake	sediment	card	change
トピック 43	105	1%	心臓	coronary	stent	cardiac	patient	artery
トピック 44	57	0%	炎症	inflammation	mouse	effect	oxidative	stress
トピック 45	31	0%	性・ホルモン	male	female	hormone	pituitary	gland
トピック 46	142	3%	金属・工学	surface	property	alloy	use	high
トピック 47	74	0%	血小板	platelet	antiplatelet	patient	dose	therapy
トピック 48	57	0%	リンパ腫	lymphoma	bcell	tcell	patient	case
トピック 49	33	6%	肝	liver	acid	metabolic	fatty	diet
トピック 50	77	0%	免疫・抗体	cell	immune	complement	response	activation
トピック 51	97	4%	画像・磁気・MRI	image	imaging	tomography	use	magnetic
トピック 52	84	1%	透析	level	serum	patient	ckd	kidney
トピック 53	63	2%	細胞・造血・再生医療	cell	stem	differentiation	human	progenitor
トピック 54	48	0%	抗凝固	patient	atrial	fibrillation	oral	anticoagulant
トピック 55	59	4%	マウス実験	mouse	use	modify	model	modified
トピック 56	141	1%	タンパク質	bind	protein	peptide	binding	acid
トピック 57	54	2%	細菌・病原体	strain	assay	bacterial	pylorus	bacterium
トピック 58	105	1%	レーザー・分析手法	use	method	laser	sample	pulse
トピック 59	80	3%	細胞・肝・増殖	cell	expression	liver	beta	pathway

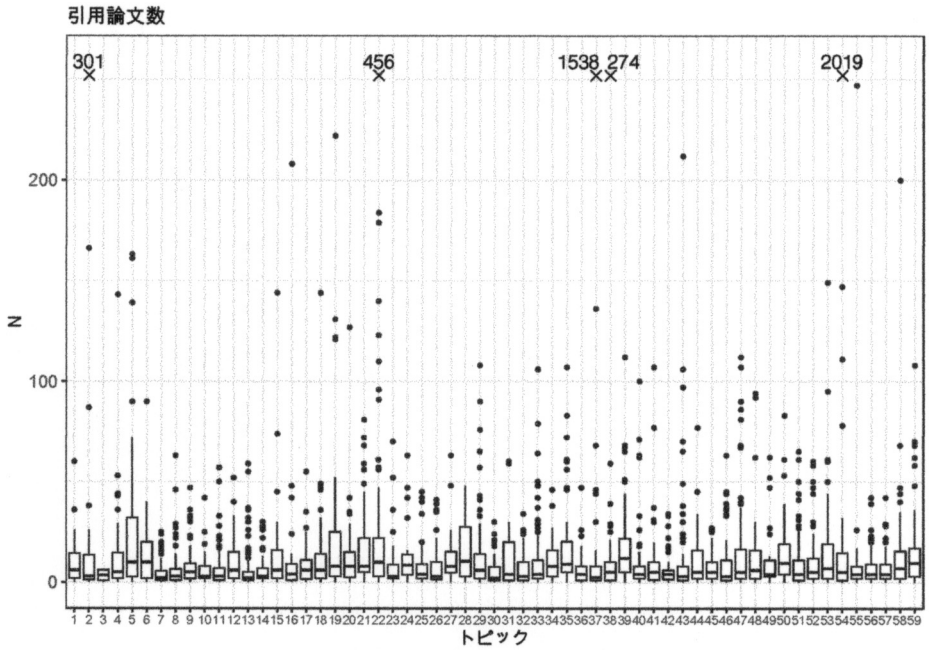

図 3. 引用論文数の箱ひげ図.

各論文のトピック出現確率のヒートマップであり，行が論文で列がトピックを示している．表 1 に 59 のトピックに振り分けられた論文数および各トピックに出現した単語上位 5 単語，それらの単語から類推される研究領域，さらに，T 大学の異なる学部の著者が含まれている論文数の割合を示した．

T 大学の場合，医学部の研究者の発表論文の割合が高いことが原因の一因だと考えられるが，推定された研究領域は，医学系分野に含まれる割合が多かった．しかし，工学系(トピック 2，46，51，58)や情報系(トピック 7)，理学系(トピック 5，42)，自然科学系(トピック 17，26，27，42)などの研究領域も推定されているため T 大学において，研究業績がある程度ある研究領域を特定することができたと考えられる．また，研究方法・手順についての記述が多いアブストラクトを用いた影響によりトピック 3 や 37 のような研究領域とは考えにくいトピックも生成された．

図 3 に各トピックに属している T 大学著者論文の被引用数の箱ひげ図を示した．外れ値により箱ひげ図が潰れてしまったため，縦軸の上限を 250 とした．被引用数 250 以上については，トピック 2 に 301，トピック 22 に 456，トピック 37 に 1,538，トピック 38 に 274，トピック 54 に 2,019 がある．また，各論文について T 大学の異なる学部の著者が含まれている割合について調べたところ(表 1 の左から 3 列目)トピック 1 とトピック 38 が 19%，トピック 10 が 10% の割合であった．トピック 10 が医学部と工学部，トピック 38 が医学部と海洋学部，生物学部，健康科学部，工学部などの共同研究であることがわかった．異なる学部の著者の含まれる論文の割合は表 1 に示した．この様にトピックモデルで研究領域を特定することで学内連携が多くみられる分野について特定することができた．

次に，T 大学引用論文の研究領域について考察する．T 大学引用論文の研究領域は，T 大学著者論文の研究成果が貢献した研究領域と考えることができる．T 大学著者論文と被引用論文

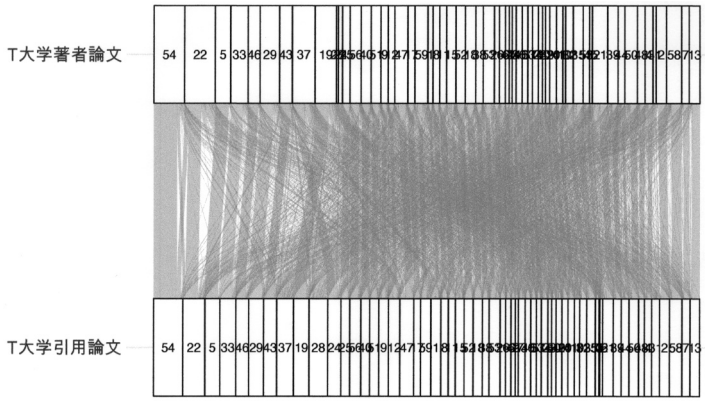

図 4. T 大学著者論文と T 大学引用論文のトピックの関連性.

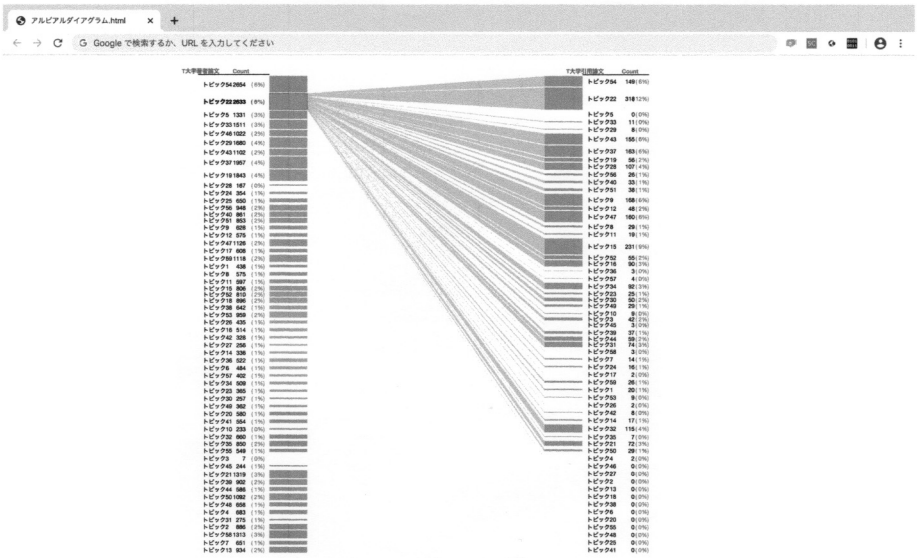

図 5. D3.js による T 大学著者論文と T 大学引用論文のトピックの関連性のインタラクティブな可視化.

の研究領域を比較することで同一研究領域または複数の研究領域での貢献であるかを把握することができる．そのために T 大学著者論文と被引用論文の研究領域の関連性をアルビアルダイアグラム(Alluvial diagram)を用いて可視化した(図 4)．図 4 の上段は被引用論文のトピック，下段が T 大学著者論文のトピックであり，それらを繋ぐ線の太さが対応する論文数を示している．図 5 は，図 4 の可視化を D3.js (Bostock et al., 2011)を用いて実現したインタラクティブな可視化である．図 5 では，左側のトピック 22 の領域にカーソルを置くことでトピック 22 に属する T 大学論文を引用している T 大学引用論文が属するトピックの連結線が強調されるインタラクティブ機能を示している．多くの T 大学著者論文は，元の論文と同じトピックに引用されているが，異なるトピックに引用されている割合が高いトピックもある．例えば，トピック 22(心臓血管)は，同一トピックだけでなく，トピック 43(心臓)，47(血小板)，54(抗凝固)な

どのトピックの論文に引用されているため，他の研究領域に影響していることがわかる．この様に T 大学引用論文の研究領域も同時に推定していたことで研究成果が貢献している研究領域についての考察も可能である．

4. 自己組織化マップによる研究実績の可視化

トピックモデルを用いて研究領域を推定することができたので，次に研究領域が類似しているトピックについて考察する．そのため推定された研究領域の類似性をクラスタリングする．図 2 では，トピックのクラスタリングが示されているが，階層型クラスター分析であるため，特定のトピックは単一のクラスターに含まれることを前提としている．トピックモデルでは各トピックの所属確率が得られるため，より詳細な分析のため所属確率を入力データとして自己組織化マップ(SOM：Self-Organizing Map)によりクラスタリングする．自己組織化マップは，Kohonen によって提案されたニューラルネットワークによりあらかじめ推定した構造にマッピングするクラスター分析の解析法でありクラスター構成を 2 次元に可視化できるのが特徴である（Kohonen, 1982, 2000）．本研究では自己組織化マップの構成は，R の kohonen パッケージ（Wehrens and Buydens, 2007)を使用した．Tian et al.（2014)では，SOM のユニット数として $5\sqrt{N}$ を目安としており，T 大学著者論文数 $N = 5,893$ について $5\sqrt{5893} \fallingdotseq 384$ であるので 20×20 の出力ユニットとした．図 6 は，20×20 の六方最密構造の自己組織化マップ上に各ユニットに振り分けられた論文のトピック番号を示し可視化した結果である．自己組織化マップでは，類似性が高いサンプルが同一ユニット及び隣接するユニットにマッピングされる．各トピック毎にトピック所属確率を標準化した値を入力データとして，自己組織化マップを適用した(図 6)．図 6 から，隣接しているトピック同士が類似性が高いトピックであるものが見受けられる．実際に下部の真ん中あたりにトピック 18(ゲノム（基礎)), トピック 19(ゲノム(応用)), トピック 10(DNA・遺伝子)とゲノムに関するトピックが隣接してマッピングされている．また，左下にはトピック 29(移植)とトピック 40(膵臓)が隣接したユニットにマッピングされている．このことから臓器移植の研究の中でも膵臓移植の研究が対象期間に T 大学において盛んに行われていることを推測することができる．

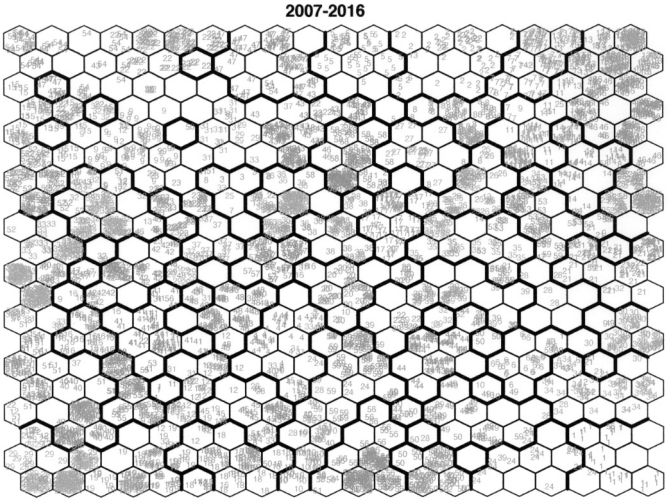

図 6. 入力データを標準化した自己組織化マップによる可視化.

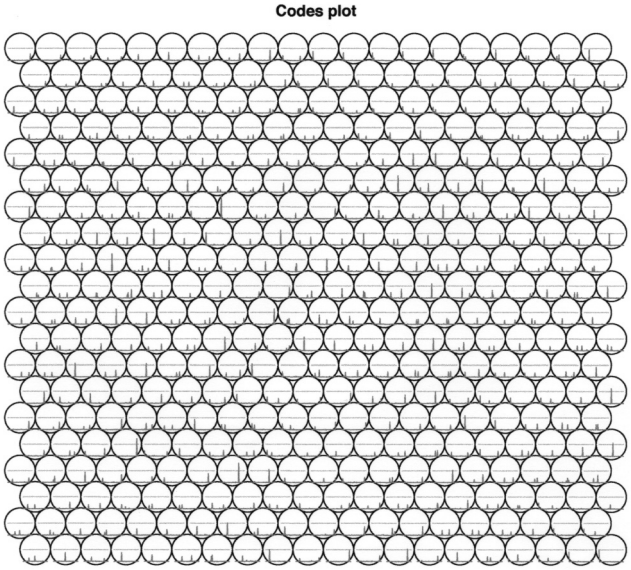

図 7. 自己組織化マップの各ユニットにおける各トピックに対する重みベクトル.

図 6 におけるユニットの境界線の太線は，各ユニットに関する自己組織化マップの重みベクトルをデータとして全 400 ユニットを Ward 法によりクラスタリングしたクラスター境界を示している．クラスター間の距離としては，ユークリッド距離を用いた．クラスター数は，Upper Tail 法 (Mojena, 1977)を用いて最適なクラスター数を算出し，37 クラスターとした．

図 6 において上から 2 段目左から 4 列目のユニットは，太線で囲まれているが，その多くはトピック 22 であり近隣のユニットにもトピック 22 がある．このユニットのトピック分布(各トピックに対する重みベクトル) (図 7)を見るとトピック 22 と同程度にトピック 34 の確率も高いことがわかった．その為，周囲のトピック 22 が多いユニットと異なるクラスターに分類された．トピックモデル数を 67 や 73 とより大きくすることでこのクラスター境界で分かれているトピックは，より小さな分野に分かれると思われる．このように自己組織化マップを用いることでトピックモデルの結果についてもより細かく分析することができる．

さらに，経年的な変化を把握するために，特定の期間に発表された論文のみをマッピングした．ここでは，2007 年から 2010 年，2010 年から 2013 年，2013 年から 2016 年の重なりのある 4 年間ごとに 3 期間に分けて可視化した．さらにトピック番号の色を各トピックに属す論文数が多いほど濃く，少ないほど淡い色とすることで，各トピックの論文数の比較ができるようにした．各期間のマップを比較することで各トピックに属する論文数の推移を捉えることができた．3 期間のマップ(図 8)を比較すると左上にマッピングされたトピック 54，トピック 47，トピック 15 の血液に関する研究領域の論文が増加していることがわかった．また，下部の中心から左側にマッピングされたトピック 18(ゲノム(基礎))，トピック 19(ゲノム(応用))のゲノムに関する研究領域は，応用研究の論文数が減少傾向にある一方，基礎研究の研究論文が増加傾向にあることがわかる．右上には，医学系以外の情報系や工学系などの研究領域が多くマッピングされており，それらは，3 期間とも論文数が多いことがわかる．

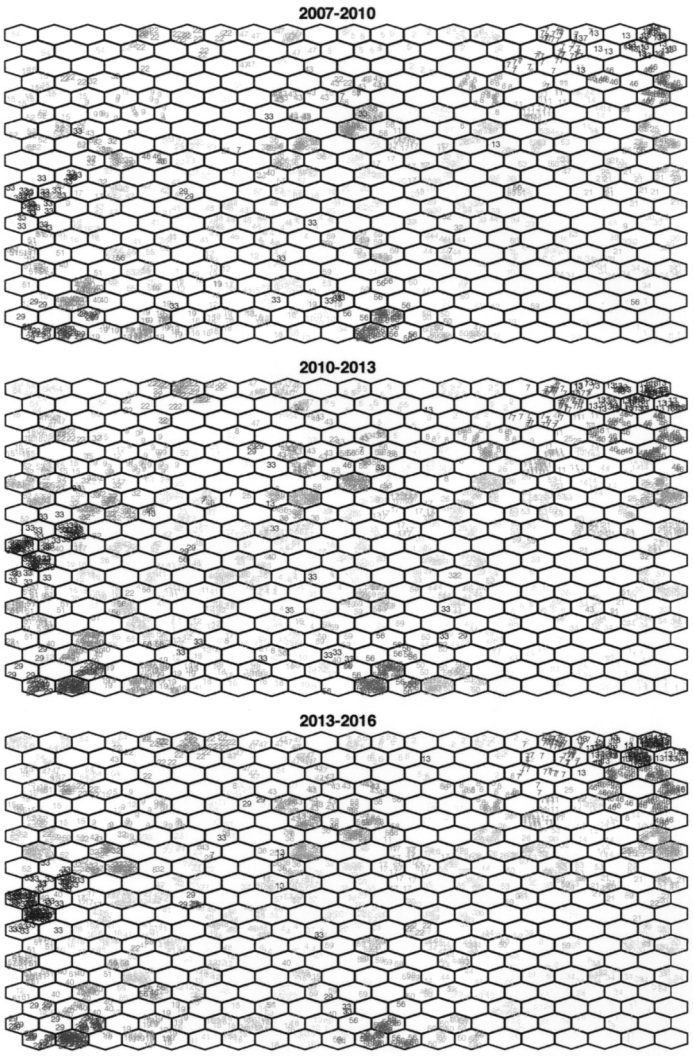

図 8. 経年的な変化の可視化.

5. 自己組織化マップを用いた可視化の改良

4 節図 6 の様に自己組織化マップを用いて，論文や研究分野を 2 次元平面上に可視化したものについて，期間を限定して表示したものを比べることで，研究が活発になった分野についても把握できることを示したが，この目的の為に有用な可視化について提案する．各トピックに属する各論文が振り分けられたユニットの中心座標をその論文の座標とし，トピックごとに重心を求め，その重心座標をトピックの代表点とする．さらに各トピックの論文数をバブルチャートを用いて自己組織化マップ上に可視化した(図 9)．この可視化により全トピックを 2 次元平面上にマッピングすることで，業績の多い研究領域について業績の量や研究領域の関連について把握しやすくなった.

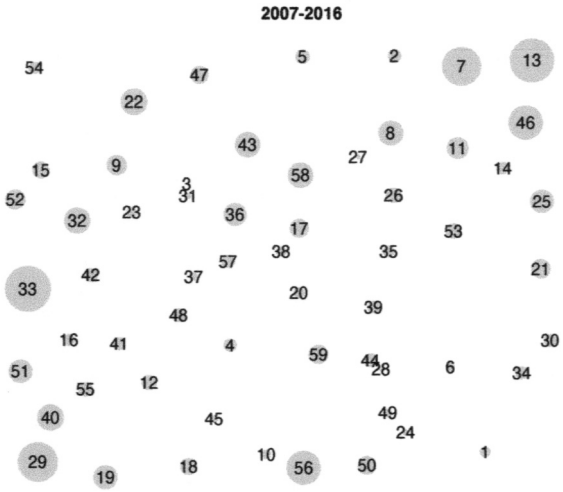

図 9. 各トピックの重心による可視化.

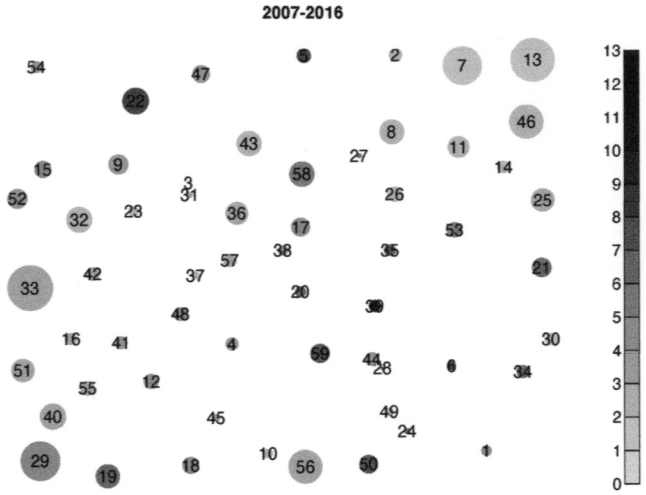

図 10. 被引用論文数の情報を追加した可視化.

研究成果は，論文数だけでなくその研究の影響力や他研究者からの評価も大切な指標である．論文の影響力は，被引用論文数で示すことができる．図 3 の被引用論文数の箱ひげ図から外れ値が多いことがわかっているので，各トピックの代表値として被引用論文数の中央値を用いた．この被引用論文数の中央値を図 9 のグラフの円の色(濃度)で示した(図 10)．被引用論文数が多いほど色を濃く，少ないほど淡くした．図 10 を見ると円が大きく色が薄い，論文数が多く被引用論文数が少ないトピックや逆に円は小さいが色が濃い，論文数は少ないが被引用論文数が多いトピックなど各研究分野について多面的に捉えることができる．前節の 3 つの期間に対する，経年変化を見ることができるようモーショングラフを作成した(図 11)．図 11 のモーショングラフは，R の plotly パッケージ (Sievert, 2018)を用いて実現しており，3 つの期間それぞれの各トピックの論文数，各トピックの論文の被引用論文数を各トピックの座標にバ

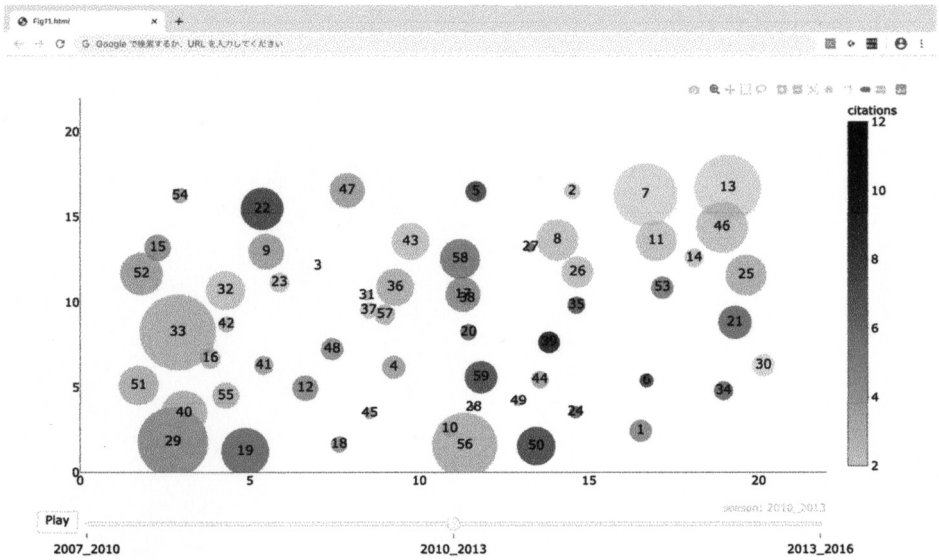

図 11. モーショングラフによる経年変化の可視化.

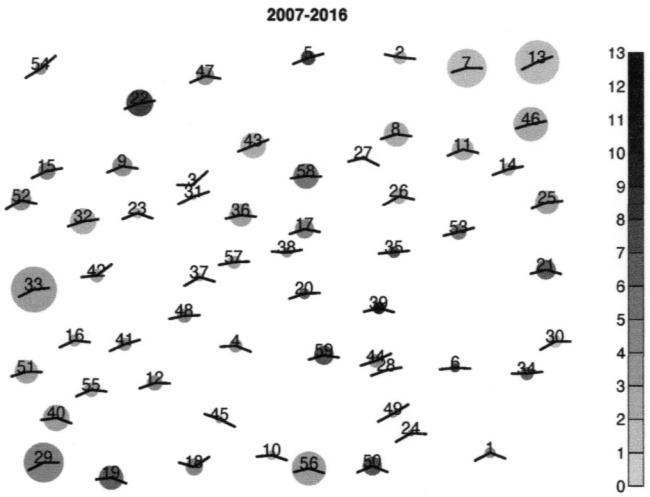

図 12. 経年的な変化の情報を追加した可視化.

ブルチャートで示し，[play] を押すことで，3 期の変化の様子が確認でき，またスライダーを動かすことで 1 期から 2 期の変化など一部の変化について確認することもできる．

さらに経年的な論文数の変化を 1 つのプロットで示す可視化を提案する．図 10 のグラフでは，各トピックの全体の論文数を円の大きさで表しているためそのトピックの論文がどの時期に増加したのかまたは減少したのかを捉えることができない．そのため図 11 のモーショングラフで見えるような論文数の推移について，3 つの期間における各トピックの論文数の増減を折れ線グラフで示し，それを図 8 の自己組織化マップ上に反映させた．図 12 は，4 節の経年的な変化の可視化と同様の 3 期間の論文数の推移を可視化した．折れ線グラフは，第 2 期(2010

年から 2013 年)を基準に前後の期間の増減を示している．この可視化により，各トピックの論文数について全体量とともに経年変化についても把握できるプロットを作成することができた．経時的な変化の可視化については，論文数だけでなく，被引用論文数の中央値，異なる学部の著者の割合などについても有用できる．T 大学の場合，自己組織化マップの右上のトピック 13(プラズマ・イオン・磁気)，トピック 46(金属・工学)，トピック 14(骨・整形外科)，トピック 25(化合物・合成)，トピック 53(細胞・造血・再生医療)の研究領域は，引用論文数はまだ少ないが論文数は経年的に増加していることが分かる．

6. おわりに

本研究では，学術文献データベースに収録されている学内所属研究者が発表した論文と，それらの論文を引用している論文のタイトルとアブストラクトのテキストデータを用いて研究領域を推定し，研究業績の多い研究領域の把握を試みた．研究領域の推定では，大規模大学においてもトピックモデルを用いることで論文の内容から研究領域を推定することができた．その際に引用している論文と一緒に分析することの有用性を示した．さらに，トピックモデルの結果を用いて自己組織化マップを適用することで，トピックモデルの結果について様々な可視化を行った．自己組織化マップによる可視化により，トピックモデルにより推定された研究領域の妥当性や類似性について把握することができることを示した．更に自己組織化マップを用いることで，推定された研究領域の業績の量や引用論文数などの可視化ができ，研究領域の関連性について把握できることを示した．また，研究領域の関連性や論文数の経時的変化を把握するための可視化についての提案を行った．

謝　辞

本研究は統計数理研究所共同利用研究重点テーマ 2「IR のための学術文献データ分析と統計的モデル研究の深化」における「学術文献 DB における著者識別問題と研究組織学術文献 DB における著者識別の精度向上に関する研究」(30-共研-4202)の助成を受けたものであり，Web of Science の DB はこの重点テーマの下で利用許可を受けている．また，本論文の執筆にあたり，有益な助言を下さった査読者の方々に心より感謝する．

参 考 文 献

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet allocation, *Journal of Machine Learning Research*, **3**, 993–1022.

Bostock, M., Ogievetsky, V. and Heer, J. (2011). D3: Data-Driven Documents, *IEEE Transactions on Visualization and Computer Graphics*, **17**(12), 2301–2309.

藤野友和, 山本由和, 船山貴光, 山本義郎 (2016). 学術文献 DB における著者識別問題について, 日本計算機統計学会第 30 回シンポジウム講演論文集, 45–48.

Grün, B. and Hornik, K. (2011). topicmodels: An R package for fitting topic models, *Journal of Statistical Software*, **40**(13), 1–30.

桂井麻里衣, 大向一輝, 武田英明 (2015). 大規模学術論文データベースにおける研究者のトピック推定と著者同定への応用, 第 7 回データ工学と情報マネジメントに関するフォーラム(DEIM2015), A5-2, 福島.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, **4**(1), 59–69.

Kohonen, T. (2000). *Self-organizing Maps*, 3rd ed., Springer, Berlin, Heidelberg.（徳高平蔵, 堀尾恵一, 大北正昭, 大藪又茂, 藤村喜久郎 訳 (2005).『自己組織化マップ』, シュプリンガー・フェアラーク東京, 東京.）

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation, *The Computer Journal*, **20**, 359–363.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees, *Proceedings of International Conference on New Methods in Language Processing*, *Manchester*.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German, *Proceedings of the ACL SIGDAT-Workshop*, Springer, Dordrecht, Dublin.

Sievert, C. (2018). plotly for R, https://plotly-r.com.

Tian, J., Azarian, M. H. and Pecht, M. (2014). Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm, *Second European Conference of the Prognostics and Health Management Society 2014*, *Nantes*.

Wehrens, R. and Buydens, LMC. (2007). Self- and super-organizing maps in R: The kohonen package, *Journal of Statistical Software*, **21**(5), 1–19.

Visualization of Research Fields Achieving Good Results in a Large University

Takamitsu Funayama[1], Yoshiro Yamamoto[2] and Tomokazu Fujino[3]

[1]Tohoku Medical Megabank Organization, Tohoku University
[2]School of Science, Tokai University
[3]International College of Arts and Sciences, Fukuoka Women's University

Large universities employ many researchers, and because research fields are extensive, it is difficult to grasp the overall research activities of a university. Understanding the research situation on a campus is necessary not only for evaluation, but also for determining future support. Therefore, in this study, we extracted text data comprising the titles and abstracts of papers contained in an academic literature database and used a topic model to estimate the research fields of those papers. In addition, we tried to estimate which research fields were achieving good results. The results demonstrated that it was possible to grasp the features of each topic classified by the topic model, as well as the relationships between topics, by visualizing the results of the topic model using a self-organizing map (SOM). We used an example to make it easy to apprehend the research trends of the university and their changes over time through the SOM visualization.

Key words: Topic model, SOM, data visualization.