統計数理 第68巻 第2号

（通巻132号）

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

目　次

特集「Institutional Research と統計科学」

2020年12月

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

〒190-8562 東京都立川市緑町 10-3　　電話 050-5533-8500（代）

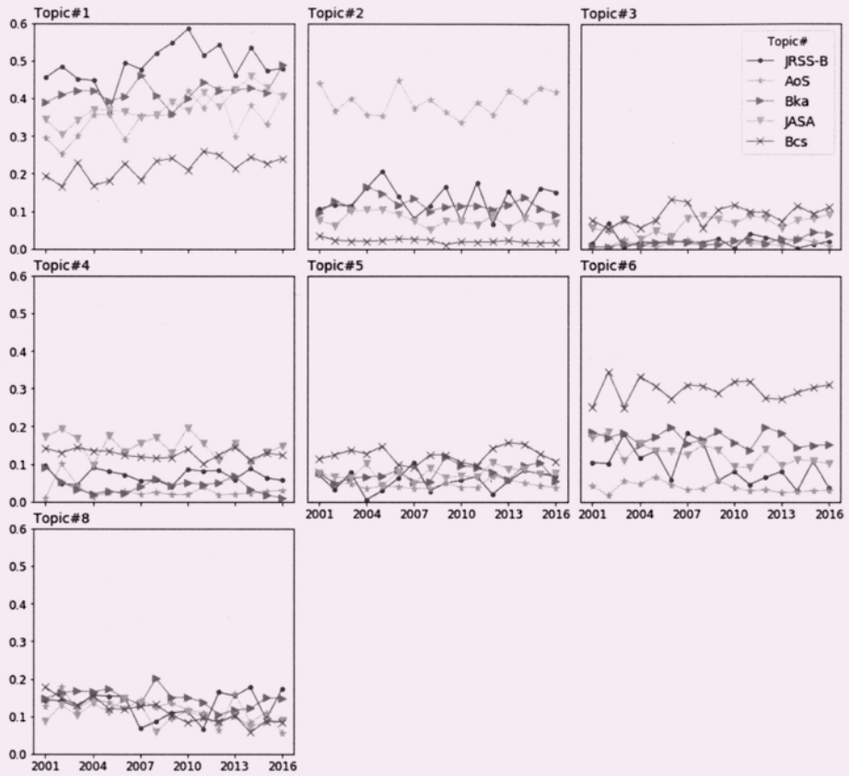
本号の内容はすべて https://www.ism.ac.jp/editsec/toukei/ からダウンロードできます

ISSN 0912-6112

統計数理　PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS　第68巻 第2号 2020

統計数理 Vol. 68, No. 2

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS



統計数理研究所

統計数理

（年 2 回発行）

「統計数理」は，統計数理研究所における研究成果を掲載する統計数理研究所「彙報」として 1953 年に歴史を始め，1985 年に誌名を変更し今の形となりました．現在は，統計数理研究所の研究活動に限らず，広く統計科学に関する投稿論文を掲載し，統計科学の深化と発展，そして統計科学を通じた社会への貢献を目指しています．

投稿を受け付けるのは，次の 6 種です．

a. 原著論文　　b. 総合報告　　c. 研究ノート
d. 研究詳解　　e. 統計ソフトウェア　　f. 研究資料

投稿された原稿は，編集委員会が選定・依頼した査読者の審査を経て，掲載の可否を決定します．投稿規程，執筆要項は，本誌最終頁をご参照ください．

また，上記以外にも統計科学に関して編集委員会が重要と認める内容について，編集委員会が原稿作成を依頼することがあります．

その他，「統計数理」に関するお問い合わせは，各編集委員にお願いします．

All communications relating to this publication should be addressed to associate editors of the Proceedings.

大学共同利用機関法人　情報・システム研究機構
統計数理研究所
〒 190-8562 東京都立川市緑町 10-3　　電話 050-5533-8500（代）
https://www.ism.ac.jp/

印刷：笹氣出版印刷株式会社

表紙の図は本誌 226 ページを参照

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

Vol. 68,　No. 2

Contents

Special Topic : Institutional Research and Statistical Science

Paper

December, 2020
Research Organization of Information and Systems
The Institute of Statistical Mathematics
10–3 Midori-cho, Tachikawa, Tokyo 190–8562, JAPAN

統計数理(2020)
第 68 巻 第 2 号 195–196

「特集 Institutional Research と統計科学」について

本多 啓介† (オーガナイザー)

この特集は統計数理研究所が実施している公募型共同利用の重点型研究「学術文献データ分析の新たな統計科学的アプローチ(2016 年度–2017 年度)」,および「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(2018 年度–2019 年度)」の 4 年間の成果をまとめたものである.この二つの重点テーマにおいては募集時に趣旨として,大規模な学術文献データを用いた多様な価値観,評価軸に基づく研究成果の分析手法や,大学・研究機関の研究活動の効果・進展を客観的に評価するための指標及び "Institutional Research"(IR)に関する方法論等について,統計科学的見地からの研究を推進することを掲げ,さらに,統計コミュニティと IR コミュニティの交流を深めるような活動を積極的に支援するとした.

国内での IR 活動の立ち上がり時期に合致した意味でタイムリーなテーマであったのか,全国の統計科学の研究者,IR に従事する大学職員,URA から多くの参加があった.累計で 51 件の課題が採択され,40 機関を超える大学・研究所に所属する研究者らが本共同利用のメンバーとして活動した.これまでの活動は ir-web という Web サイト (本多・濵田, 2016)にまとめられている.この重点テーマに採択された課題はトムソン・ロイター社(現クラリベイト・アナリティクス社)の大規模書誌データベースである Web of Science Core Collection が研究目的のため利用できることが大きな特徴であった (統計数理研究所プレスリリース, 2015).

IR という言葉には多義性があり,特に国内の多様な導入状況においてはその概念を一義的に定めることは困難 (小林・山田, 2016)であるが,ここでは書誌データや研究業績データ,その他学内データ等を対象とした分析手法の開発を行い,機関の意思決定に利活用されることを志向するものをおおまかではあるが「研究 IR」とする.特に個々の機関での意思決定に資する分析手法の確立を目指す,という点がビブリオメトリクス(計量書誌学)やジャーナルインパクトファクター等の単純な利用との大きな違いと考えている.

本特集「Institutional Researchと統計科学」では,特に大規模書誌データを利用した研究 IR に関する 6 篇の論文から構成されている.その内訳は原著論文が 3 篇,研究ノートが 1 篇,総合報告が 2 篇である.以下簡単に紹介する.

山田論文は国内の高等教育における IR の進展の経緯と現状が詳細にまとめられている.著者は米国発祥の IR に対して日本での経営 IR・教学 IR の概念 (小林・山田, 2016)を導入したオーソリティの一人であるが,本論文では筆者らの国内大学を対象としたアンケート調査を元に URA 整備事業が研究 IR を推進している動向を分析した.筆者はここに日本の高等教育特有の動向があると指摘している.藤野・濵田論文は,著者識別を論文要旨の情報から同定する手法として複数のトピックモデルのアルゴリズム比較を行っている.武井・藤野・中野論文は,同じくトピックモデルの一種を用いて,ある集団の研究動向を時系列に把握するための手法を提案している.船山・山本・藤野論文はこれもトピックモデルで大規模大学における研究

† 統計数理研究所:〒 190–8562 東京都立川市緑町 10–3

者群の研究領域の特定を行った上で，さらに自己組織化マップを組み合わせて領域間の関連性を可視化した．以上 3 篇は書誌データに対して大規模な自然言語処理を適用したアプローチと言える．張・潘・中野論文は学術分野の関係性の分析として統計科学関連とその他分野の引用関係を調べ，分野間の相互作用の度合いを数値化した．水上・中野論文は異分野融合の進展を見るために分野間の繋がりを共著関係に注目して，IoT 分野の研究論文の国別比較を例に分析を行っている．この 2 篇は書誌データの引用-被引用，共著というネットワーク構造から研究動向を捉える手法の提案である．

4 年間実施された重点テーマの目的の一つであった統計科学コミュニティと IR コミュニティの交流推進には大きな成果があった．重点テーマ参画のグループを中心に日本計算機統計学会のスタディグループ「IR(Institutional Research)のための統計的モデル構築に関する研究」が設置され，国内の大学評価担当者の大学間連携の組織である大学評価コンソーシアムとの交流を開始したことである．これまで両機関と統計数理研究所が共催で事務系職員を対象とした統計の基礎を学ぶ勉強会として「初歩的な統計講座」を複数回実施してきた．いずれの回も参加申し込みがすぐに埋まるほど盛況であった．受講者の満足度も高く，今後はさらに IR の実務に即したテキストの充実などの交流を通した効果が期待できる．

2016 年度からの 4 年間の公募実施期間は各大学等において IR 室の設置や活動が活発化した期間と重なるといってよい．国立大学等の第 3 期中期目標・中期計画の策定と開始のタイミングと重なったこともあるだろうし，実際この時期，統計関連の研究者から，「ちょうど所属の大学でも IR を担当するようになった」という声をよく耳にした．あらゆる現象を客観的に評価する，データに基づく意思決定を行う，必要であればそのための指標を開発する，というのは言うまでもなく統計科学の領域であるが，IR はまさに灯台下暗し，当たり前のように研究者が所属する大学や研究機関周辺にこそ新たな開拓領域があると言えないだろうか．この特集で多くの研究者が面白さを発見するとともに，各機関内で研究者，IR 実務者，URA が密接に連携した活動が今後も活発になることを期待する．

最後に，この特集「Institutional Research と統計科学」の査読者の方々，編集担当の方々，並びにクラリベイト・アナリティクス社に，この場をお借りして感謝を申し上げたい．

参 考 文 献

本多啓介, 濵田ひろか (2016). 統計数理研究所が取り組む IR 機能強化, https://ura3.c.ism.ac.jp/ir-web/.
小林雅之, 山田礼子 (2016). 大学の IR 意思決定支援のための情報収集と分析, 慶應義塾大学出版会, 東京.
統計数理研究所プレスリリース (2015).「統計数理研究所とトムソン・ロイターが協力体制を構築」, https://www.ism.ac.jp/ura/press/ISM2015-02.html.

統計数理(2020)
第 68 巻 第 2 号 197–208

特集「Institutional Research と統計科学」
［総合報告］

日本における IR の動向：経営 IR，教学 IR から研究 IR の誕生と推移

山田 礼子†

（受付 2019 年 9 月 4 日；採択 2020 年 1 月 15 日）

要　　旨

米国の大学で 1960 年代から発展してきた IR への関心が近年日本でも高まっている．米国における IR は大学の経営支援，意思決定支援，戦略計画，教学改善とアセスメントといった領域では定着しているが，研究に関する IR は，機関内の IR 部門ではなく，研究担当部門が行うなど分散型システムが基本である．一方日本では，グローバル大学を多くの大学が目指し，ランキング上昇という必要性かつ外的な要因から「研究 IR」に従事する大学が増加している．日本における IR は政策動向にあわせて変化し，その多様な機能から，IR の概念を統一することは極めて困難でもある．本稿では，第一に，IR の定義の多義性を検討しつつ，日本の高等教育を巡る環境との関係から IR の動向を検討する．次に，日本における研究 IR を本研究では，「掲載ジャーナルの質(質的指標)，論文数や被引用数(量的指標)の測定等を行い，機関としての研究力向上に資する活動」と定義し，研究 IR に携わっている URA を対象にしたウェブ調査を通じて，研究 IR による研究成果の可視化が URA を活用している大学のどの層の大学に影響を与えているのかを検討する．調査結果からは，研究 IR が，主に大規模研究大学だけではなく，1 件あたりの採択金額の小さい大学ほど改善効果があるという知見が得られた．

キーワード：IR，研究 IR，URA 調査，研究成果の可視化，大学ランキング．

1. はじめに

近年，大学における IR についての関心が高まっている．大学機関調査や機関研究と訳されることも多い IR は米国の高等教育機関で 1960 年代に誕生したといわれている．小林・山田(2016)は，IR 部門は，教育，経営，財務情報を含む大学内部の様々なデータの入手や分析と管理，戦略計画の策定，アクレディテーション機関への報告書や自己評価書の作成を主な仕事として，米国の多くの高等教育機関に常設されており，こうした活動から，組織運営に関する意思決定の支援部門というニュアンスが強い一方で，教育改善のためのデータを集積，分析し，改善に活かすという教学 IR も IR の重要な役割であり，学内の教育の質保証にも深く関与しているのが IR 部門とみなしている．日本でも高等教育の質保証推進政策を背景として，GPA 制度，CAP 制の導入，単位の実質化等の方策が既に多くの大学で実施されるようになった．さらには，2012 年の中央審議会による「審議まとめ」では，質を伴った学修時間の実質的な増加・確保を始点とした好循環という視点からの学士課程教育の充実が求められたことを契機として，既に導入されている教育改善のための方策を十分に機能させ，教育の質保証を推進するために

† 同志社大学 社会学部：〒 602–8580 京都市上京区烏丸今出川東入る

は，IR と呼ばれる機能の開発や部門の設置がより強く求められるようになった．教育情報の公表に伴い，データを一元化し，様々なデータベースに情報を提供するだけでなく，情報を検索して報告書を作成していくために加工することも IR 部門の新たな仕事として認識されている．また，大学のガバナンスの整備が求められるなかで，ガバナンスの支援機能としての IR という見方も浮上しつつある（山田, 2016a）．

このように近年，日本の大学においても IR 活動の実践への取り組みが散見されるようになっているが，現在の日本における IR は政策動向にあわせて変化し，その多様な機能から，IR の概念を統一することは極めて困難であるという現実も否定できない．大学の規模，大学の特質，設置形態によっても IR が目的とするところは様々であるからである．IR の定義については，その最初の創設国である米国においても様々であり，実践活動も多様である．日本においては，高等教育政策の流れのなかで，IR 部門の設置が国立大学法人および私立大学にも求められているが，IR そのものの定義や活動は一致していない．本稿では，第一に，こうした IR の定義の多義性を検討しながら，日本の高等教育をめぐる環境との関係から IR の動向との関連性を検討する．その際，IR 発生の地と認識されている米国での研究に関する IR の状況について焦点を当てる．第二に，日本における研究 IR の推進が大学に及ぼす効果について URA 調査を基に考察する．

2. IR の定義の多様性

IR は Institutional Research の略語であるが，機関研究と文字通り翻訳してもその意味は捉えにくい．小湊・佐藤（2012）は「機関の計画策定，政策形成，そして意思決定を支援する情報を提供するために，高等教育機関内で行われる調査研究」とするサウプの定義（Saupe, 1990）が最も広く受け入れられていると論じている．実際に，様々な論者が IR の定義について議論しており，サウプの定義にもとづいた IR の活動も，データを収集して調査報告することから始まり，学習成果[1]（中央教育審議会, 2012）のアセスメント，全学レベルの戦略計画の策定など多様であるのみならず，その活用の範囲も大学によって幅広い（山田, 2016a）．

IR の活動を推進するうえで，「情報」は重要な概念である．Fincher（1978）は，IR を「組織的情報力（organizational intelligence）」と定義しているが，この定義をさらに発展させ組織的情報力を 3 層構造から分析したのが Terenzini である．Terenzini（1993）は，第 1 層をファクトブックや統計情報の作成に不可欠な基礎的な「技術的・分析的情報力」，第 2 層を技術に関する知識に加えて高等教育に特有の課題に関する知識を持ちかつ課題の解決に向けて取り組むことができる「問題に関する情報力」，第 3 層は自大学という内部の状況を外部環境という文脈から相対的に捉え，分析できる「文脈的情報力」として構造的に説明した．Terenzini の 3 層構造から成る組織の情報力という概念は，IR の組織体制，機能，そして IR 担当者の専門職としての発展を説明する概念として使用されることになる（山田, 2016a）．

浅野 他（2014）は，先人の IR を巡る議論を踏まえて，IR には様々な定義・用法があり，一義的には定まっていないとみなし，その理由として IR が各大学の活動として散発的に発展してきたことにあることをあげている．また，IR 活動の内容が機関の性格によって異なること，つまり，ワールド・クラスの大学であるか，研究中心の大学であるか，地域志向性が高い大学であるか，そしてより限定的なコミュニティに根差す志向性のある大学であるか，あるいは公立か私学であるかという設置形態によって目指す IR 活動が異なることが指摘されている（Delaney, 2009; Leimer and Terkla, 2009）．

しかし，IR 活動が機関の性格，機関の規模，設置形態等によって多様であったとしても，IR 活動が何のために行われるのかというその目的と役割という点から定義するとすれば，異なる

側面が見えてくるのではないか．情報やデータの収集，分析，情報やデータを報告すること，それらのデータや情報の解釈の手助けを行うことは「決定支援」としてみなされており，Webber (2018) は今日の高等教育環境においては，IR は機関の意思決定に貢献する活動であるとし，機関の意思決定に貢献する IR 活動は，米国を含む北アメリカ，ヨーロッパ，ラテンアメリカ，南アフリカそしてアジア諸国の IR 活動の共通要素であるとしている．換言すれば，IR の明確な定義や用法が定まっていないにせよ，大学の経営に関する意思決定，教育の改善，さらには戦略計画策定のために，大学内外に存在するデータを収集し，クリーニングをしたうえで，分析し，活用することが IR の基本原理であり，この基本原理は国境を越えて広く認知されているとみなされる．

3. 日本における IR の進捗状況と特徴

IR における基本原理は，国境を越えても「機関の意思決定に貢献する活動であり，そのために情報の収集，分析，解釈を行うこと」[2] であるといえるが，その領域は，大学経営，教学，そして研究に関することにまたがっている．こうした IR の原理と活動領域を視野にいれたうえで，日本の IR の進捗状況と特徴はどのようなところにあるのかを調査をもとに把握する．

2014 年に国公私立大学 783 校を対象に実施した文部科学省先導的大学改革推進委託事業「大学における IR の現状とあり方に関する調査研究」（東京大学）の調査結果を参照すると，IR 組織の設置状況について，「IR 名称の組織がある」(9.9%) と「IR 名称はないが，担当組織がある」(15.1%) は合わせて約四分の一となっている．「全学レベルの組織がない」割合は 69.1% を占めているが，IR 組織を設置していない大学のうち，設置に関して，「検討中」が 36.1% となっており，IR 組織の設置が計画のなかに組み込まれつつあることがうかがえる．IR 組織の設置目的（複数回答可）としては，「教育改革の成果のチェック」，「大学評価への対応」が 6 割を超えており，「大学経営上の必要性」は 57.1%，「学生への支援」(48.1%)，「大学の説明責任を果たすため」(38.5%) が重要な設置目的の上位を占めている．IR 組織の担当業務については（複数回答可），「執行部への調査情報・分析の提供」(65.6%)，「認証評価への対応」，「文部科学省の大学政策のウオッチ」が 5 割強，「大学改革動向のウオッチ」が 5 割弱と上位を占め，「学生による授業評価の分析」，「学生の達成度調査」等教学に関する調査・分析も 40% 前後となっている．一方で，財務に関する分析や 10% 未満と低い数値を示している．調査結果からは，①IR 組織はガバナンスとの連関から設置され，ガバナンスへの貢献を視野に入れて，執行部への情報の提供・分析を行う比率が高くなっているという政策動向に応じて IR の変化の兆しが見られること，②評価対応および情報への対応は外部との連関から重要視されつつあること，③全般的に学習成果対応の教学 IR が推進されるなど，教育の質保証への対応が進捗しているといったことが確認されている（山田, 2016b）．

日本の中で教学 IR が各大学において推進されてきていることとして，先ほどの財務関係の IR がそれほど進んでいないことと比較した場合，財務や施設に関するデータは，個別の大学の内部情報として外部に明らかにしにくい性格のものが含まれているだけでなく，他の大学と共有しにくい性格も伴っている．一方，教育に関する学生のデータ，例えば，学生調査は個別の大学のみならず，多くの大学が共通して利用できるだけでなく，結果を教育の効果に関するベンチマークとして利用することも可能である．こうした教学 IR にみられる特徴から，教学 IR が日本の大学で進捗している状況に関係していると捉えられる．

4. グローバル化の影響と研究 IR の登場

IR の基本原理と日本の IR の進捗状況を調査から検討してきたが，近年の外的要因としてグ

ローバル化と大学ランキングによる影響により，新たな課題が浮上している．

Marginson and Roades (2002)は，グローバル化による経済的，社会的，知識上，教育上，そして文化的な影響をいずれの高等教育機関も受けると主張した．この主張を所与のものとして，Botha (2018)は，グローバル化が IR にもたらすインパクトには，留学生や外国人教員の流動化の促進，外国大学との協定の増加，海外企業との連携の増加等により，信頼できるデータを作成し，情報を活用することが含まれ，結果として，IR 部門にそうした活動をより強く求めるようになると論じている．グローバル化と知識社会との関連では，ワールド・クラスの研究大学もしくはワールド・クラスを目指す研究大学にとってはグローバルな知識ネットワークの形成が活発化する（Altbach et al., 2010）．同時に，国境を越えての研究成果の競争も激化することで，とりわけ研究大学の執行部は，研究者に関する情報，公表されている論文数，本の出版数，会議の proceedings 数等の研究成果に関するデータ，大学院生数やポスト博士号研究者情報，研究資金獲得情報，産学連携契約数，知財契約数等を含む研究に関する総合情報が研究マネジメントを実施していくうえでの重要な情報と認識し，どの研究分野に注力していくか，共同研究をいかに推進していくか，ベンチマーキングをしていくための該当機関の選定，そしてランキング情報とその改善のためにどうすべきか等戦略的研究マネジメントとして位置づけている（Botha, 2018）．しかし，Botha (2018)はこうした研究に関する情報やデータとマネジメントに関する研究マネジメント機能は米国では IR 部門の役割ではなく，研究支援部門の役割であるとみなしたうえで，縦割り的サイロにならないように，他の IR 部門との情報交換が不可欠であると主張している．一方で，米国の研究志向大学の数は限られていることから，IR 部門にとって研究マネジメントは大きな関心ごとではないことにも言及している．

研究 IR の推進の前提条件として，論文数(質的指標)，被引用数(量的指標)の測定については，世界的にみても合意は存在しているが，筆者が長らく IR 研究に携わってきたなかで，「研究 IR という概念の形成」が本当に広く認識されているのかについては疑問があるだけでなく，Botha が言及しているような状況もある．そこで，こうした状況を前提に IR が誕生し，IR の活動が多面的に行われている米国の状況のインタビュー調査結果を提示する．米国の IR 研究者と実践者が多く参加する Association for Institutional Research(AIR)幹部と研究者にインタビュー調査を 2016 年に実施し[3)]，米国に研究 IR という概念が存在するのか，世界の大学ランキングとの関係，ワールド・クラス大学というブランディングとの関係性について主に調査をした．インタビューを通じて得た知見をもとに米国における研究に関する IR とそれを取り巻く状況についてまとめてみる．

組織上，研究担当副学長と教学担当副学長は別ラインである．この組織的特性は日本の大学機関でも同様である．米国では，研究 IR はどこが，どのように担当し，推進しているのかについては，研究に関しては，分散システムが基本となっていることから，各スクール，各学科が研究評価に責任を持っているところが多いということである．したがって，大学内で中央集権的な部門として機能している IR 部門は，研究に関する IR を担当することは少ないとのことであった．

大学ランキングと IR の関係性については，多くの米国の大学が意識するランキングは，「US News & World Report ランキング」，「ベストカレッジ・ランキング」であり，これらのランキングの指標のほとんどが教育に関する内容から構成されている．ランキングを作成するプロセスは，①研究型大学(修士・博士課程を持つ研究中心全国型大学)，②全国型リベラルアーツ大学(学士課程中心大学)，③地域型総合大学(学士課程中心，少数の修士・博士課程)，④地域型リベラルアーツ大学(地域に根差したリベラルアーツを中心とした学士課程大学)といった大学の機能分類にもとづいて，教育に関する指標が基本となる学術的な質に関する量的な手法でランキングが決定される．

例えば，研究型大学と全国型リベラルアーツ大学に適用されるベストカレッジに関する 15 の指標例[4] では，卒業率と一年次生残留率（合計で 22.5%），peer 大学機関の執行部による評価（22.5%），教員のリソース（教員数，教員給与，教員の取得学位，フルタイム比率，クラスサイズ，ST 比）（20%），学生の選抜度（SAT/ACT 点数，合格率）（12.5%），学生対費用（学生一人当たり教育費用）（10%），graduation rate performance（学生の教育的付加価値，テスト点数等）（7.5%），卒業生の大学への順位付け（5%）高校の進路カウンセラーによる順位付け等が重要視されているなど，国内でのランキングが求める指標の多くが教育に関する内容つまり academic excellence である．

大学院レベルで重要視されているランキングでは，「ベスト Graduate Schools ランキング」であり，このランキングでは分野別プログラムが単位の基本となっている．大学院プログラムやグローバル大学ランキングにおいては，教員リソースに関して教員の研究力が重要な指標となるものの，やはりプログラムの内容，教育に関すること，労働市場に関することに焦点がおかれている．ただし，教員のサラリー分析などは国内大学ランキング指標にも挙げられていることから IR 担当者が分析することは多いとのことであった．

日本での研究 IR 担当者の仕事として Web of Science やスコーパス等を利用して論文に関する生産性や被引用数のデータ分析が主なものであるのに対し，多くの米国大学の IR 担当者は，こうした分析には携わっていないという．米国における IR は大学の経営支援，意思決定支援，戦略計画，教学改善とアセスメントといった領域では定着しているが，研究については米国の多くの研究志向の大学には研究担当の副学長のもとに研究に関連する分野を扱う部門が置かれ，そこで研究力の向上施策や支援を行っていることが一般的である．それゆえ，IR 担当部門は，経営 IR の観点から教員の生産性等の分析には関わるが，研究の側面に IR 部門が関わることはほとんどないといっても過言ではない．それでは，研究 IR は日本型 IR の特徴であろうか．こうした問題意識に立ちながら，次節では日本の「研究 IR」について検証する．

5. 日本型 IR の特徴である研究 IR と先行研究の動向

グローバル化の影響に伴い，ワールド・クラス大学を目指す研究志向大学においては，研究のマネジメントは重要であるとの認識はされているものの，米国における研究に関する IR は，研究担当部門の役割として認識され，実際に IR 部門の活動もしくは役割として位置づけられていないこと，また研究志向大学以外の米国の大学の多くは教育の充実を指標としている国内ランキングに対処するべく，教学 IR を充実させる傾向があることをインタビュー調査から確認した．

一方日本においては，研究志向型大学に限らず，グローバル大学への転換を多くの大学が目指し，ランキングを上昇させる必要性という外的な要因から IR が関与するかもしれない新たな領域として，高等教育機関では「研究 IR」への関心が高くなっているように見える．日本においては，高等教育政策の流れのなかで，IR 部門の設置が国立大学法人および私立大学にも求められるようになってきているが，IR そのものの定義や活動は一致していないことは先述した．まして，研究 IR についての定義や活動，研究 IR 人材に関する能力・スキル要件，専門職コミュニティの組織構造，要件についての議論も緒についていないと言える．一方で，実態としては，文部科学省が平成 23 年度より URA（ユニバーシティ・リサーチ・アドミニストレーター）整備事業として，各大学等で研究開発に知見のある人材を URA として活用・育成する制度を整備している．文科省の支援を背景として，URA 部門を設置し，URA を置く大学，研究機関等も増加している．URA は，大学教員の科研申請書の支援から，研究プロジェクトの企画立案，機関としての研究力向上に関係する研究 IR 分析などを行うなどの幅広い活動を行っ

ている．

こうした日本の現状に鑑み，日本における研究 IR を本研究では，研究 IR とは「掲載ジャーナルの質（質的指標），論文数や被引用数（量的指標）の測定等を行い，機関としての研究力向上に資する活動」と定義した上で，研究 IR に携わっている職種である URA を対象にしたウェブ調査を通じて，研究 IR による研究成果の可視化が URA を活用している大学のどの層の大学に影響を与えているのかを検討する．

研究 IR が重要な課題になっている環境においては，研究 IR として開発した評価指標を参考に，資金・要員の調達や基盤整備といった資源再配置から，成果を導くことが期待される．それゆえ，評価指標の多くは，論文数や被引用数（量的指標）の測定研究に重点が置かれてきた．統計数理研究所は，2016 年度の重点テーマとして，「学術文献データ分析の新たな統計科学的アプローチ」を採用し，トムソン・ロイターの協力のもと，IR でも重要なテーマとなっている研究機関・大学の研究成果分析の手法や研究活動の進展，効果を客観的に評価するための指標，および IR に関する方法論等について，統計科学的見地からの研究を推進していくことを発表したが，学術文献データ分析を通じての書誌情報分析が研究 IR の研究の方向性でもある．例えば，同研究所の助成を受けた濵田 他（2016）による「大学研究力評価のための書誌情報の分析」では，学術文献に現れるデータは共著関係や引用―被引用関係として示すことができるというネットワーク構造を持っているとしたうえで，研究の発信状況，研究の協力状況，研究の影響力引用関係を分野別およびある大学と他機関との協力関係を検証している．この研究からは研究者間の分野あるいは異分野，他機関とのネットワーク関係を把握することができる．他にも，多視点で即時的な影響度を波及効果として測定する手法が，林・山下（2011）による「ビブリオメトリクスを用いた大学研究活動の自己分析」，吉田（2014）による「計量書誌学の新たな挑戦」，荒木 他（2017）による「大学における部局横断型共同研究の活発さを把握する指標の検討」，そして豊田（2019）による包括的な日本の研究力についての実証的研究により考察されている．

これらの研究は評価指標の開発として位置づけられるが，一方で研究 IR による研究成果の可視化が機関としての大学にどのような影響を与えてきているのか，さらには，米国とは異なり研究志向大学に限らず，多くの大学がワールド・クラス大学への転換を意識しているように見受けられる状況において，大学のどの層に影響を与えてきたのかという問いには対応してはいない．本研究は，URA（B 調査）と URA 部門あるいは研究部門に責任を持つ部門長あるいは理事（副学長）を対象にした調査（A 調査）のうち B 調査結果からこの問いをベースに検討するという差異がある．

6. 調査の概要

IR は機関の計画策定，政策形成，そして意思決定を支援する情報を提供するために，高等教育機関内で行われる調査研究として認識され，黎明期にあった日本の IR は学習成果に関連した教学 IR および国立大学を中心とした評価対応としての IR を軸に展開してきた．先行の IR 調査結果からは，私立大学に限定すれば，沖（2013）による私学高等教育研究所 IR 調査や日本私立大学連盟（2018）による IR 調査等では比較的大規模私立大学が，積極的に IR 組織を設立してきたことが提示されている．一方，研究 IR データの活用は研究中心の大規模国立大学において進んでいると推察されるが，実際には研究 IR の推進状況や改善の度合いはどのようになっているのかを把握することが日本の研究 IR の進展具合とその効果を検証することにも貢献できると考えられる．

URA 整備事業は比較的大規模な大学に役立つ施策であったのかを検証することを目的として調査を設計し，平成 29 年 12 月に全国の国公私立の URA を対象にメールでの連絡による

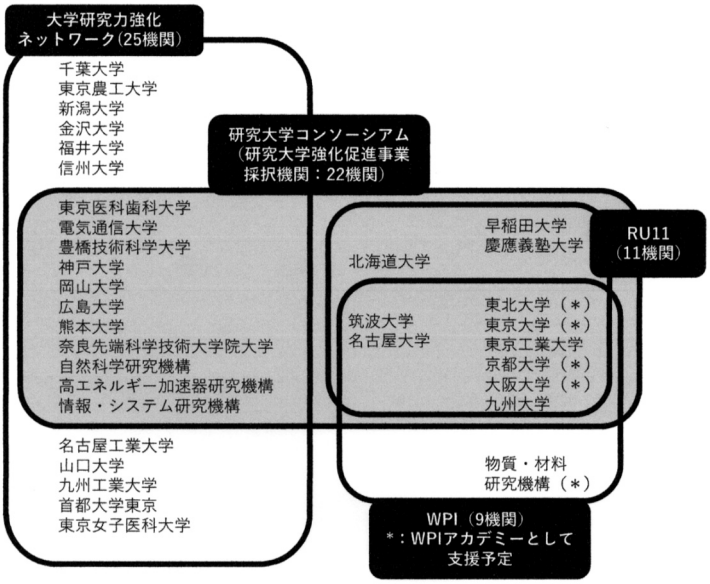


図 1．URA 部門を設置している主な配布先の例．「研究大学強化促進事業推進委員会(第 7 回，平成 29 年 5 月 25 日開催)資料 3−2 研究大学のネットワーク化について」より筆者作成．https://www.mext.go.jp/a_menu/kagaku/sokushinhi/1386456.htm

ウェブ調査による実査を行った．主な配布先は，図 1 に示す「大学研究力強化ネットワーク」加盟 25 機関，「研究大学コンソーシアム」加盟 22 機関，「RU11」に属する 11 機関，「WPI」9 機関の合計 34 機関に加えて，20 機関は URA 協議会を通じて連絡をした．重複している機関があるため図 1 に示した機関からの回答数は 12 であり，URA 協議会を通じての回答数は連絡数通りの 20 機関となり，合計 32 機関であった．

有効回答数 43(回答機関数 32)，内訳は国立大学 22，公立大学 1，私立大学 8，その他 1 であった．回答機関を 2017 年 10 月公表の科研費データから機関の採択件数で分類すると，250 件未満 10 件，250 件以上 500 件未満 19 件，500 件以上 1000 件未満 8 件，1000 件以上 6 件であった．1 件当たりの採択金額では，2000 千円以下 16 件，2000 千円以上 2500 千円未満 11 件，2500 千円以上 3000 千円未満 7 件，3000 千円以上 9 件となった．有効回答数の少なさから今回の調査データについて多変量解析等を実施することは不可能であることから，全体像を把握するために単純集計によるクロス分析を中心に提示する．

6.1　調査結果の概要

回答者の属性は，研究活動活性化のための環境整備及び研究開発マネジメント強化といった URA 業務の担当(1)が 17 名(39.5%)，組織や教育研究等の情報収集や分析をして，意思決定・評価改善活動支援を行う IR 業務(2)を担当している者が 1 名(2.3%)，上記 1(URA 業務)と 2(IR 業務)を兼務している者が 23 名(53.5%)，その他 2 名(4.7%)という結果となった．

URA 業務経験年数は平均 5.1 年(SD: 3.8)，URA 業務経験大学数は平均 1.3 校(SD: 0.72)である．比較的新しい「専門職」として，大学内で雇用されてきていること，一方で，URA として大学間を移動している数はそれほど多くないことから，URA の労働市場が日本において展開していくかどうかは現時点では判断できないとみなされる．

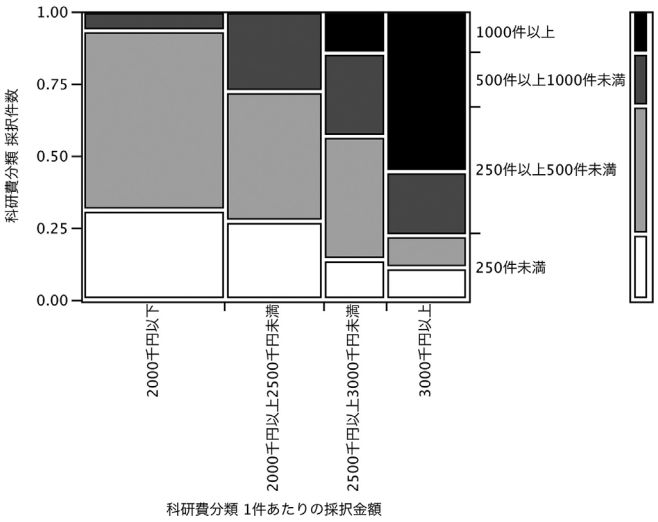


図 2. 1 件当たり科研費平均と科研の採択件数の関係.

採択件数と 1 件当たりの科研費平均のクロスを図 2 に示しているが，間接経費を含めて 1 件あたりの科研費平均が 200 万以下の大学は，地方国立大学や私立大学が多いこと，採択件数 1000 件以上，科研費平均 300 万円以上の大学は研究大学強化促進事業採択機関が多いという特徴が見られた.

URA(ユニバーシティ・リサーチ・アドミニストレーター)整備事業が研究 IR を推進してきている背景があるとすれば，URA の所属組織，そして IR 部門との関係を探ることも必要である．所属する大学にどのような組織が設置されているかについては，全学の IR 組織 58.1%，部局の IR 組織 2.3%，全学の URA 組織 23.3%，部局の URA 組織 4.6% となっているが，一方で「設置予定があるが現時点で未設置」が 7.0%，「設置予定無し」が 14% であった．回答者の所属先については，全学の IR 組織 7.0%，部局の IR 組織 0，全学の URA 組織 65.1%，部局の URA 組織 7.0%，全学の知財組織，研究推進部，産学連携組織などのその他 20.9% となっていた．URA 組織に属し研究 IR 活動に従事している回答者が多いことが読み取れる．業務内容については，研究活動活性化のための環境整備及び研究開発マネジメント強化といった URA 業務を担当している割合が 39.5%，組織や教育研究等の情報収集や分析をして，意思決定・評価改善活動支援を行う IR 業務を担当が 2.3%，上記に定義される URA 業務と IR 業務を兼務して担当している割合が最も多く 53.5% となっている．URA 業務と IR 業務との兼務が多いという回答から，URA と IR 担当者という名称や所属組織は別に存在しているものの，米国とは異なり，研究マネジメントと IR 業務において，大学内での役割分担がなされていない特徴的な構造が浮かび上がっている．こうした回答傾向から，日本型研究 IR の特徴の一つに，経営 IR や教学 IR，そして研究 IR が組織の意思決定に貢献する活動と位置づけられているために，包括的な IR 活動として機能することが期待されているのではないかとも考えられる.

改善案を提案するという研究 IR の貢献について見てみると，1 件あたり「採択金額少」の大学で貢献が高い傾向が見られ，データ分析は「採択件数」の少ない大学は積極的ではない傾向がある．採択件数の少ない大学は，データの提供にも積極的ではない傾向が確認された.

研究者個人の業績評価については，採択件数，採択金額の多寡にかかわらず，いずれの大学も研究者個人の業績評価に積極的な傾向が見られる一方で，部局全体の業績評価については，

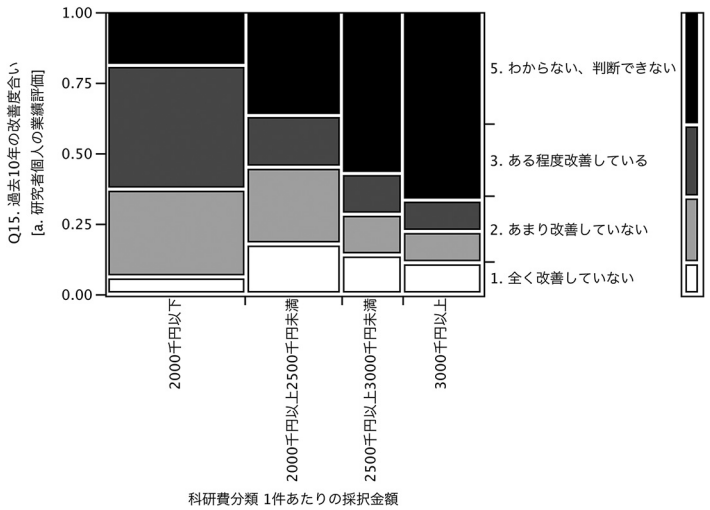


図 3. 過去 10 年の 1 件当たり科研費平均と研究者個人の業績評価の関係.

採択件数の少ない大学，1 件あたりの採択金額の小さい大学に消極的傾向が散見された．特に，科研費獲得は採択件数の少ない，1 件あたりの採択金額が小さい大学ほど，消極的傾向が強い．受託研究費獲得や寄付金獲得，さらには研究者のマッチングを行うことで新しい分野の開拓につながる活動については，採択件数，1 件あたりの採択金額が大きいほど積極的である．このことから，研究 IR は，採択件数や 1 件あたりの採択金額の大きい大学ほど，積極的であるが，「研究者個人の業績評価」については，1 件あたりの採択金額の小さい大学も積極的に実施しているといえよう．

それでは過去 10 年間での研究に関するマネジメントの改善についてはどうだろうか．過去 10 年の改善度合いについて，研究者個人の業績評価の改善度合いについてみると図 3 に示しているように 1 件あたり採択金額の少ない大学ほど，過去 10 年で改善したと回答する傾向が高い．一方，部局全体の業績評価は，採択件数の多い大学ほど，過去 10 年に改善したと回答する傾向が見られた．科研費獲得，寄付金獲得については 1 件あたりの採択金額の小さな大学ほど，過去 10 年で改善したと回答する傾向が高い．直近だけでなく，10 年間スパンで見ると，研究 IR は，1 件あたりの採択金額の小さい大学に，改善効果を与えたことが限定データから確認できた．

7. おわりに

IR 先進国米国における研究マネジメントと経営 IR や教学 IR が同じ部門ではなく，研究マネジメント部門で行われているという分散型システムであるのに対し，日本では URA という職種が登場しているものの，通常の IR 業務との兼任を行っているなど研究 IR も経営 IR や教学 IR と同じ組織の意思決定に資する活動を行う部門という位置づけがなされ，それらを総合的に IR 活動としているところに日本型 IR の特徴が見られ，同時に研究 IR という用語が定着しつつあることも日本の動向であるといえよう．かなり，米国の IR 組織及び活動の発展の軌跡とは異なる様相で進捗しているとも捉えられる．しかし，URA 組織を大学内で設置し，URA を置いていることそのものが，今回の調査対象校の少ないことが示しているように，研究 IR を進捗させようとしている大学数が多いとはいえないことの証左でもある．それにもかかわら

ず，大規模研究大学以外の中堅大学においても，研究 IR を進捗させていこうとする姿勢が確認できたことは，日本型研究 IR の構造であるとも見て取れる．

研究 IR と機関の研究マネジメントへの貢献という視点で分析すると，特にこの 10 年間というスパンで見た場合，研究 IR 全般において，1 件あたりの科研費採択金額の小さい大学ほど，過去 10 年で改善傾向が確認された．このことは，どちらかというと教学 IR が分析部門の専門性を持つ人材不足から専門性を持つ IR 担当者とその部門を設置するだけの余裕がある大規模国立・私立大学を中心に展開してきた様相とは異なる様相を示している可能性がある．すなわち，研究 IR が，主に，大規模研究大学のものとしてのみ発展してきているわけではなく，むしろ，1 件あたりの科研費採択金額の小さい大学ほど改善効果があるというところに特徴があるのではないだろうか．今回は，限定的なデータであるが，研究 IR やリサーチアドミニストレーター，URA のいわゆる「措置効果」がどの層の大学にみられるかを，今後慎重に精査していく必要があると考えられる．

注．

[1] 『予測困難な時代において生涯学び続け，主体的に考える力を育成する大学へ』（中教審審議まとめ）から，それまで使用されていた「学習」が大学設置基準に合わせて「学修」として標記されるようになっているが，ここでは米国で使われている．Learning Outcomes の一般的な翻訳が「学習成果」であることから，「学習」あるいは「学習成果」として標記する．

[2] IR の多義性の議論を基にまとめた共通要素である．

[3] 2016 年 5 月には AIR の幹部である Clifford Adelman 博士にインタビューを行い，10 月には IR 研究者でもあるノースカロライナ州立大学の Stephen Porter 教授と Paul Umbach 教授へのインタビュー調査を実施した．

[4] 本稿では 15 の指標すべてを例示してはいない．

参 考 文 献

Altbach, P., Reisberg, L. and Rumbley, L. E. (2010). Tracking a global academic revolution, *Change*, **42**(2), 30–39.

荒木将貴, 桂井麻里衣, 大向一輝, 武田英明 (2017). 大学における部局横断型共同研究の活発さを把握する指標の検討, 日本データベース学会和文論文誌, **16**(1), 1–6.

浅野茂, 黄文哲, 小林雅之, 森利枝, 山田礼子, 劉文君 (2014). 平成 24–25 年度文部科学省大学改革推進委託事業,『大学における IR の現状と在り方に関する調査研究報告書』, 東京大学, http://www.mext.go.jp/a_menu/koutou/itaku/__icsFiles/afieldfile/2014/06/10/1347631_01.pdf, http://www.mext.go.jp/a_menu/koutou/itaku/__icsFiles/afieldfile/2014/06/10/1347631_02.pdf.

Botha, J. (2018). The impact of global forces in higher education on the development of institutional research, *Building Capacity in Institutional Research and Decision Support in Higher Education* (ed. K. L. Webber), 19–36, Springer International Publishing, Cham, Switzerland.

中央教育審議会 (2012). 予測困難な時代において生涯学び続け，主体的に考える力を育成する大学へ（審議まとめ）, 文部科学省, 東京.

Delaney, A. M. (2009). Institutional researchers' expanding roles: Policy, planning, program evaluation, assessment, and new research methodologies, *New Directions for Institutional Research*, **143**, 29–41.

Fincher, C. (1978). Institutional research as organizational intelligence, *Research in Higher Education*, **8**(2), 189–192.

濵田ひろか, 森裕一, 飯塚誠也, 本多啓介 (2016). 大学研究力評価のための書誌情報の分析, 日本計算機統計学会 第 30 回シンポジウム講演論文集, 43–44.

林隆之, 山下泰弘 (2011). ビブリオメトリクスを用いた大学研究活動の自己分析, 情報管理, **53**(12), 665–679.

小林雅之, 山田礼子 (2016).『大学の IR：意思決定支援のための情報収集と分析』, 慶応義塾大学出版会, 東京.

小湊卓夫, 佐藤仁 (2012). 米国における IR 機能発展の背景,『IR 実践ハンドブック：大学の意思決定支援』(大学評価・学位授与機構 IR 研究会 編), 玉川大学出版部, 東京.

Leimer, C. and Terkla, D. G. (2009). Laying the foundation: Institutional research office organization, staffing and career development, *New Directions for Institutional Research*, **143**, 43–58.

Marginson, S. and Rhoades, G. (2002). Beyond national states, markets and systems of higher education: A glonacal agency heuristic, *Higher Education*, **43**(3), 281–309.

日本私立大学連盟 (2018).『これまでの IR・これからの：課題と提言』, 日本私立大学連盟, 東京.

沖清豪 (2013). 私立大学における IR の現在, 2013 年調査の結果から, アルカディア学報, No.557, 教育学術新聞, 平成 26 年 5 月 28 日.

Saupe, J. L. (1990). *The Functions of Institutional Research*, 2nd ed., Association for Institutional Research, Tallahassee, Florida.

Terenzini, P. T. (1993). On the nature of institutional research and the knowledge and skills it requires, *Research in Higher Education*, **34**(1), 1–10.

豊田長康 (2019).『科学立国の危機』, 東洋経済新報社, 東京.

Webber, K. L. (2018). Institutional research and decision support in higher education: Considerations for today and for tomorrow, *Building Capacity in Institutional Research and Decision Support in Higher Education* (ed. K. L. Webber), 3–18, Springer International Publishing, Cham, Switzerland.

山田礼子 (2016a). アメリカにおける IR の展開―IR 機能に伴う二面性と専門性を中心に―, 高等教育研究, 第19集, 25–48.

山田礼子 (2016b). 日本の IR の現段階, IDE 現代の高等教育, **586**, 11–16.

吉田光男 (2014). 計量書誌学の新たな挑戦：国産オルトメトリクス計測サービスの開発, 情報の科学と技術, **64**(12), 501–507.

Trend of IR in Japan: Emergence of IR Including Management, Teaching and Learning and Research

Reiko Yamada

Faculty of Social Studies, Doshisha University

In recent years, interest in IR, which was developed at universities in the United States in the 1960s, has increased. IRs in the US are well established in areas such as university management support, decision support, strategic planning, academic improvement and assessment, but IRs related to research are decentralized, and are conducted by the department in charge of research rather than within the organization's IR division. On the other hand, in Japan, many universities aim to achieve a global reach, and the number of universities engaged in "Research IR" is increasing due to external pressure to raise their ranking. The IR in Japan currently changes with policy trends, and it is extremely difficult to provide a unified definition of IR because of its various functions. This paper first examines the definition of IR and explores the relationship with the trend of IR from the relationship with the environment surrounding higher education in Japan. Next, after defining research IR so as to measure the quality of published journals (qualitative indicators), the number of articles and the number of citations (quantitative indicators), etc., through a web survey targeting URA, we examine what type of university is affected by the visualization of research results of research IR activities. Based on the results of the web survey, the research IR might have had a beneficial effect not only in the immediate past, but also over a 10-year span, on universities that accepted a small amount of money per case of scientific research expenses. Despite the limited data at this time, research IR is not necessarily developed for only large-scale research universities, but also has a positive effect on universities with a small amount of money accepted per case of scientific research expenses.

Key words: Decision making support, IR, research support, academic excellence, ranking, quality assurance.

統計数理(2020)
第 68 巻 第 2 号 209–218

特集「Institutional Research と統計科学」
[総合報告]

学術文献DBにおける著者識別のための トピックモデリングの利用とその性能比較

藤野 友和[1]・濵田 ひろか[2]

(受付 2019 年 7 月 23 日;改訂 11 月 19 日;採択 12 月 4 日)

要　　旨

本研究では,学術文献データベースから,特定の組織に所属する研究者が著者となっている論文リストを,統計的自然言語処理の一手法であるトピックモデリングを用いて抽出する方法を提案する.当該組織名が所属に含まれている論文に対してトピックモデリングを適用し,著者ごとの特徴ベクトルを作成する.これに基づいて,当該組織の研究者の名前のみがマッチして,組織名が含まれていない論文について著者識別を行う.この識別性能を,複数のトピックモデル(Latent Dirichlet Allocation,Dirichlet Multinomial Regression,Correlated Topic Model)で比較した.

キーワード:学術文献データベース,研究力評価,統計的自然言語処理,Institutional Research.

1. はじめに

大学や研究機関が自組織の研究力の評価を行う際に基本となるデータは,所属する研究者の研究業績リストである.この情報を研究者の申告によって収集することも可能であるが,所属する研究者全員に提出を依頼したり形式を統一したりするなどの作業コストがかかるうえ,業績の多い研究者であるほど,提出されたリストに漏れが出てくる可能性も高まる.そこで,この作業を自動化することが望まれる.Web of Science(WoS)や SCOPUS などの学術文献データベースには主要な雑誌や会議録に収録された論文に関する情報が収められており,ここから論文リストを抽出することが可能である.データベースの提供者が,研究者の ID を整備し,データベースに追加された新規の論文について,研究者 ID との紐付けを正しく行っていれば,ある組織に所属する研究者の ID リストを用いることで,必要とする情報を抽出できる.しかしながら,Tang and Walsh (2010)は,WoS などの主要な学術文献データベースにおいても,研究者の ID 付けは完全でなく,完全にある研究者を特定するには至っていないと指摘している.また,データベースには同姓同名の研究者も非常に多く含まれており,これが著者の識別を困難にしている大きな要因となっている.したがって,本研究では,データベースで付与された著者 ID には頼らずに著者識別を実施する方法を提案する.データベースで付与された著者 ID を用いて,つまり信頼して,著者の特徴づけを行った上で,新たな識別対象論文の著者同定を行う手法が桂井 他 (2015)で提案されている.桂井 他 (2015)は,日本の学術文献データベー

[1] 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
[2] 統計数理研究所 特任研究員:〒 190–8562 東京都立川市緑町 10–3

スである CiNii を対象に，トピックモデリングを用いた著者識別の方法を提案した．この方法は，以下のような手順によるものである．

(1) データベースからすべての著者の論文を最大で 5 本ずつ収集し，それらのアブストラクトを学習データとする．
(2) 学習データに対してトピックモデリングを適用し，著者ごとのトピックの特徴ベクトルを算出．
(3) 判定対象論文の特徴ベクトルと著者の特徴ベクトルの類似度を計算し，類似度が最大の著者を選出．

本研究では，データベース内の文献のうち，著者の所属に自組織の大学名や研究機関名が含まれており，かつ，現在自組織に所属している研究者が著者に含まれているもののみを学習データとして用いる．そして，識別対象論文が自組織の著者によるものであるかどうかの判別を行う．これに対し，桂井 他 (2015) は，識別対象論文の著者名と同姓同名の著者の候補が複数おり，そこから特徴ベクトルの類似度が最大となる著者を選出するという最適化の問題であり，本研究とは問題設定が異なる．また，著者の特徴づけのために桂井 他 (2015) と同様にトピックモデリングを用いるが，LDA (Latent Dirichlet Allocation) だけでなく，複数の拡張手法を利用して識別精度の向上を試み，性能比較を行った結果について考察を与える．

なお，著者識別には，本研究で提案するような論文の内容に対する自然言語処理をベースとするものだけではなく，共著情報や引用，被引用情報などを利用したネットワーク分析などを利用する方法も考えられる．本研究の位置づけは，自然言語処理の枠組みでどの程度まで識別が可能かどうかを探るものであり，最終的にはこれらの組み合わせによって，高精度の識別ができるようになることが望ましい．

2. トピックモデリング

トピックモデルは，統計的自然言語処理の主要なモデル群のひとつであり，文書の単語に潜在的なトピックを仮定するモデルの総称である．Blei (2012) により提案された LDA は，トピックモデルの中でも基本的なモデルであり，自然言語処理において広く用いられている．LDA では，以下のような文書の生成モデルを仮定する．

1. 各トピック t に対して，$\boldsymbol{\phi}_t \sim \mathrm{Dirichlet}(\boldsymbol{\beta})$
2. 各文書 d に対して
   (1) $\boldsymbol{\theta}_d \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$
   (2) 各単語 i に対して
      (i) $z_{d,i} \sim \mathrm{Multi}(1, \boldsymbol{\theta}_d)$
      (ii) $w_{d,i} \sim \mathrm{Multi}(1, \boldsymbol{\phi}_{z_{d,i}})$

$\boldsymbol{\phi}_t$ はトピック t ごとの単語の出現分布を示す，大きさがデータセット全体の語彙数 V のベクトルである．$\boldsymbol{\theta}_d$ は各文書の全単語における潜在トピックの出現分布を示す，大きさがトピック数 K のベクトルである．各単語のトピック $z_{d,i}$ が，パラメータ $\boldsymbol{\theta}_d$ のカテゴリカル分布，すなわち大きさ 1 の多項分布から生成され，そのトピックに対応した $\boldsymbol{\phi}_{z_{d,i}}$ をパラメータとするカテゴリカル分布から単語 $w_{d,i}$ が生成される．LDA のグラフィカルモデルによる表現を図 1 に示す．

一方，Mimno and McCallum (2008) は，文書ごとのトピック分布を生成するパラメータ $\boldsymbol{\alpha}$ に，回帰構造を導入した Dirichlet Multinomial Regression (DMR) トピックモデルを提案した．

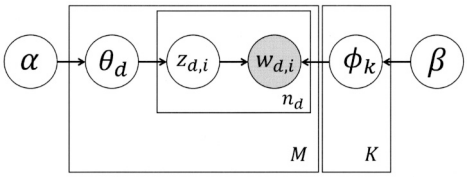


図 1. LDA のグラフィカルモデル表現.

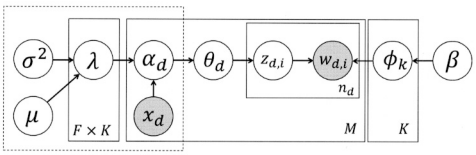


図 2. DMR トピックモデルのグラフィカルモデル表現.

これにより，文書に付随する情報をトピック分布に反映させることができ，より安定的なトピック分布を得ることができるようになる．DMR トピックモデルの生成モデルは以下のようになる．

1. 各トピック t に対して
   (1) $\boldsymbol{\lambda}_t \sim N(0, \sigma^2 I)$
   (2) $\boldsymbol{\phi}_t \sim \mathrm{Dirichlet}(\boldsymbol{\beta})$
2. 各ドキュメント d に対して
   (1) 各トピック t に対して $\alpha_{dt} = \exp(\boldsymbol{x}_d^T \lambda_t)$
   (2) $\boldsymbol{\theta_d} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_d)$
   (3) 各単語 i に対して
      (i) $z_{d,i} \sim \mathrm{Multi}(1, \boldsymbol{\theta}_d)$
      (ii) $w_{d,i} \sim \mathrm{Multi}(1, \boldsymbol{\phi}_{z_{d,i}})$

また，グラフィカルモデルによる表現を図 2 に示す．本研究では，説明変数として著者の組織内における所属学科などの情報を用いる．

以上のモデルでは，トピック間の相関は考慮されないが，学術文献から抽出されるトピック間には相関があるとする方が自然である．例えば，統計関連のトピックと数学関連のトピック，英語関連のトピックと文学関連のトピックなどの組み合わせはある程度相関があると考えられる．Lafferty and Blei (2006)は，このようなトピック間の相関を考慮したモデルである Correlated Topic Model(CTM)を提案した．LDA の生成過程において，トピックの出現確率 $\boldsymbol{\theta}_d$ はパラメーター $\boldsymbol{\alpha}$ のディリクレ分布に従う部分の代わりに，多変量正規分布 $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ に従う $\boldsymbol{\eta}_d$ を導入して，単体上のベクトルを得るために以下の変換によって $\boldsymbol{\theta}_d$ を得る．

$$\theta_{d,t} = \frac{\exp(\eta_{d,t})}{\sum_{t'=1}^{K} \exp(\eta_{d,t'})}, \quad t = 1, \ldots, K$$

これによって，トピック間の相関を表現することができる．CTM のグラフィカルモデルによる表現を図 3 に示す．

本研究では，以上 3 つのモデルについて著者識別の性能比較を行う．LDA と CTM については統計解析ソフトウェア R の topicmodels パッケージ(Grün and Hornik, 2011)を用いてモデ

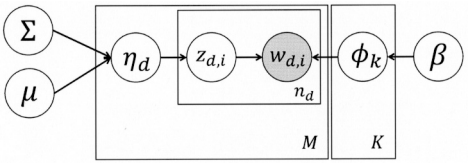


図 3. CTM のグラフィカルモデル表現.

ルの学習を行った．また，DMR トピックモデルの学習については Java で開発された Mallet (McCallum, 2002)を用いた．モデルの学習アルゴリズムは，LDA と DMR はギブスサンプリング，CTM については，変分ベイズ法によるものである．性能比較を実行する目的においては，アルゴリズムを統一する方が望ましいかもしれないが，DMR トピックモデルを実装している主要なソフトウェアパッケージは現在のところ Mallet のみであるため，このような形となった．また，著者情報からトピックを生成するモデルである著者トピックモデル(Rosen-Zvi et al., 2004)は DMR トピックモデルの特別な場合と考えることができる．一方，結合トピックモデル(Lin and He, 2009)は文書につけられた補助情報のトピックをトピック分布から生成し，そのトピック上の補助情報分布から補助情報を生成するモデルである．この補助情報を当該組織の著者かどうかを示すフラグとして，これを予測するタスクとして著者識別問題を捉えることもできる．しかしながら，本研究の問題設定では，後述のように学習データとして当該組織の著者による論文のみを利用するため，この手法を採用することはできない．

3. 提案手法の概要

本研究で提案する著者識別手法の概要は以下のとおりである．

(1) 自組織の研究者リストを準備する．
(2) 研究者リストに基づいて，学術文献データベースより著者名が研究者名に一致する論文のアブストラクトをすべて抽出する．
(3) そのうち，著者名の所属が自組織の名称と一致するものを学習データとする．自組織の名称が含まれない論文を，識別対象論文とする．
(4) 自組織に所属する研究者の研究紹介文を学習データに加える．
(5) 学習データに対し，10 分割交差検証を実施し，最適なトピック数を検討する．
(6) 学習データに前の手順で得られたトピック数を用いてトピックモデリングを適用し，著者ごとのトピックの特徴ベクトルを算出する．
(7) 識別対象論文に対して，著者の特徴ベクトルに基づく確信度を算出し，著者識別を実行する．

(4)において，研究者の研究紹介文を学習データに加えているのは，自組織に着任して間もない研究者や，学術文献データベースに収録されている雑誌の少ない分野の研究者の学習データが不足するのを補うためである．

最適なトピック数については，手順(5)において，10 分割交差検証によって決定する．本研究では，各文書の単語を学習用とテスト用に分割する方法ではなく，学習用文書とテスト用文書に分割する方法を採用した(佐藤, 2015)．各文書の単語を分割する方法では，分割をした場合に十分なサイズの単語数が確保できない．それは，学術文献データベースに論文データが存在しない研究者の文書は，研究紹介文のみとなり，論文データが存在する研究者に比べて単語数が少なくなってしまうためである．研究紹介文の単語数も研究者によってばらつきが大きく，

単語数が少ない場合はこの問題が顕著となる．学習用とテスト用の文書集合をそれぞれ $\boldsymbol{d}^{train}$, $\boldsymbol{d}^{test}$ とし，さらにテスト用の文書集合における文書に含まれる単語を $\{\boldsymbol{w}_d^{test1}\, \boldsymbol{w}_d^{test2}\}_{d=1}^{M^{test}}$ に分割する．$\boldsymbol{d}^{train}$ に含まれる単語集合 $\{\boldsymbol{w}_d^{train}\}_{d=1}^{M^{train}}$ と $\{\boldsymbol{w}_d^{test1}\}_{d=1}^{M^{test}}$ を用いてモデルを学習し，学習アルゴリズムに応じて $\{\boldsymbol{w}_d^{test2}\}_{d=1}^{M^{test}}$ に対する対数尤度 $\mathcal{L}^{test2}$ を計算する．さらに，以下の式で与えられる Perplexity

$$\mathrm{PPL}^{test2} = \exp\left\{-\frac{\mathcal{L}^{test2}}{\sum_{d=1}^{M^{test}} n_d^{test}}\right\}$$

を求める．Perplexity はモデルの汎化性能を示す指標であり，小さいほど汎化性能が高いことを意味する．Perplexity が十分小さくなるトピック数で得られるモデルを著者識別に用いる．

各モデルの学習で得られた著者 d のトピックの出現割合の事後分布 $\hat{\boldsymbol{\theta}}_d$ および，トピック k における単語の出現割合の事後分布 $\hat{\boldsymbol{\phi}}_k$ を用いて，著者 d の文書において単語 w_i が出現する確率を

$$p(w_i|d) = \sum_{k=1}^{K} \hat{\theta}_{d,k}\hat{\phi}_{k,i}$$

と推定する．著者識別を行う際には，著者 d によるものかどうかの識別対象となっている論文 d' に出現するすべての単語について $p(w_i|d)$ を計算する．論文 d' が著者 d によるものであれば，その著者の専門分野に関する特徴的な単語の $p(w_i|d)$ は大きくなる傾向があると予想される．それと同時に，多くの一般的な単語や，新たにその論文で扱われる専門用語などの $p(w_i|d)$ の値は小さくなり，その出現頻度は相対的に多い．これらのことを考慮して，論文 d' が著者 d によって書かれたものであるかどうかの確信度を

$$\mathrm{conf}(d'|d) = F^{-1}(0.75)$$

により定義する．ただし，関数 F は文書 d' に対する $p(w_i|d)$ の経験分布関数である．

4. 対象データ

本研究では，第 1 著者の所属先である福岡女子大学に所属する研究者の氏名，研究紹介文および論文の情報を分析対象とした．研究者情報は 2017 年 3 月時点のものであり，対象者は 87 名である．研究紹介文は，公式ウェブサイト（`http://www.fwu.ac.jp/`）に掲載されている研究者情報より，スクレイピングで英語で書かれた研究紹介文を取り込んだ．日本語のみを掲載している研究者については，日本語の研究紹介文を取り込んで機械翻訳によって英文に変換したものを利用した．学術文献データベースは，クラリベイト・アナリティクス社から WoS の収録データの提供を受けて独自に整備したグラフデータベースを用いた．グラフデータベースはオープンソースソフトウェアの Neo4j（Neo4j, 2019）を用いている．WoS の収録データの期間は 2005 年から 2014 年までの 10 年間であり，ここから福岡女子大学に所属する研究者の氏名に基づいた検索を行い，2422 件の論文を抽出した．このうち，所属情報に福岡女子大学が含まれているものは 249 件あり，これを学習データとして用い，残りの論文が識別対象となる．

アブストラクトや研究紹介文については，レンマ化およびストップワードの除去を行ったものを利用した．レンマ化は単語の見出語を取得して品詞を特定するための処理であり，例えば estimates，estimated や estimating を estimate に揃える目的で用いられる．ストップワードの除去については，R の tidytext パッケージ（Silge and Robinson, 2016）に収録されているストップワード辞書に収録されている 728 単語に一致する単語をすべて除去した．また，研究に関する記述によく見られる単語，例えば study や research などについては，逆文書頻度を計算して，

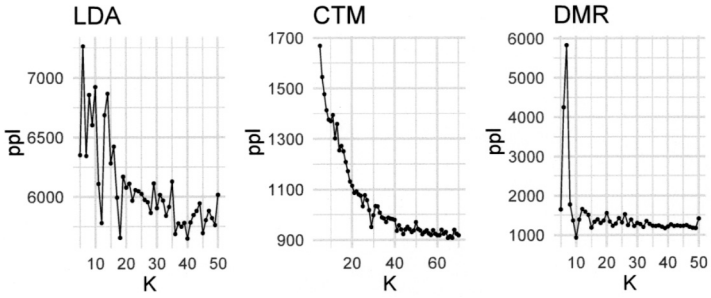


図 4. 各モデルに対する Perplexity.

表 1. 判定結果.

	当該組織の研究者の論文	当該組織の研究者の論文でない
当該組織の研究者の論文と予測	a	b
当該組織の研究者の論文でないと予測	c	d

値が 1 未満のものを除去した．単語 w に対する逆文書頻度 $\mathrm{idf}(w)$ は以下の式で与えられる．

$$\mathrm{idf}(w) = \log \frac{|M|}{|\{d \in M | w \in d\}|}$$

ただし，M は文書全体の集合で，$|M|$ は全文書数を表す．これにより，study や research のほか，develop や analysis のような単語も除去された．

5. 適用結果および考察

図 4 に，10 分割交差検証によって得られたトピック数と Perplexity のプロットを示す．分割ごとに得られる Perplexity については平均値を採用している．LDA では，Perplexity はトピック数が 40 付近で下げ止まり，以降はトピック数を増やしても 5500 から 6000 の付近で推移する．CTM では，トピック数が 55 付近で Perplexity が下げ止まり，以降は 900 を下回らない程度で推移する．DMR においては，トピック数が 30 から 40 付近まで緩やかに減少し，以降は 1000 から 1500 の間で推移する．Perplexity で見ると，LDA よりも CTM や DMR の方がよい汎化能力を持っていることがわかる．

次に，学習データ全体でモデルを学習し，これによって得られた著者ごとのトピック分布とトピック上の単語分布を用いて，識別対象論文の確信度を計算した．得られた確信度を昇順にソートし，当該組織の研究者が執筆した論文かどうかを判定する閾値をその確信度に設定した場合に，どの程度よく判別できるかどうかを F 値を計算することによって確認した．判定結果が表 1 となった場合，F 値は適合率 $a/(a+b)$ と再現率 $a/(a+c)$ の調和平均で定義される．

トピック数に対して最大の F 値をプロットしたものを図 5 に示す．各モデルにおいて，最大の F 値の最大値およびそのトピック数は，LDA が $0.58(K = 31)$，CTM が $0.64(K = 54)$，DMR が $0.64(K = 41)$であった．図 4 と図 5 を比較すると，Perplexity が十分下がったトピック数におけるモデルを使うことで，高い F 値が得られることが確認できる．F 値以外にも，break-even ポイントや 11 点平均精度も計算したが同様の傾向を示した．F 値自体は Perplexity と同様に，LDA よりも CTM や DMR の方が良好な値を示した．

図 6 は，CTM においてトピック数が 54 の場合の，論文ごとの確信度の対数をプロットした

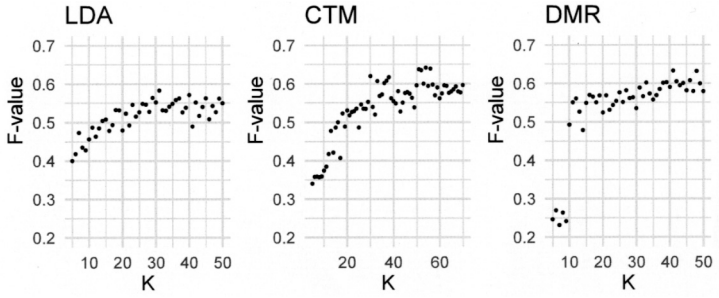


図 5. 各モデルに対する最大の F 値.

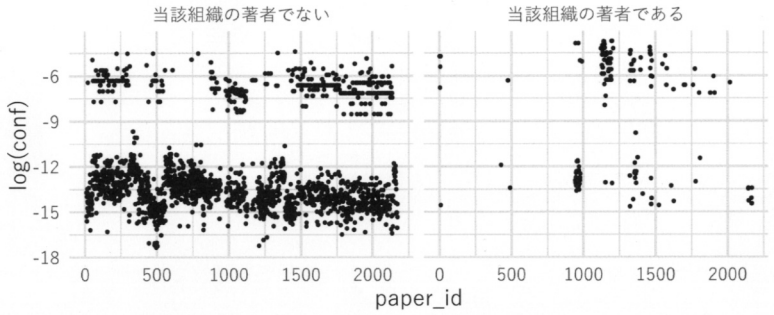


図 6. トピック数を 54 とした CTM による確信度の分布.

もので，識別対象論文において，当該組織の著者による論文とそうでないものに分割して示している．いずれの場合も確信度の対数が -9 となる付近で分布が 2 つに分かれている．この確信度を閾値として識別を行った場合，当該組織の著者であると判定された論文で誤判別されたものと，正しく判別されたもので確信度の対数の平均値を比較すると，前者が -6.73 ± 0.67 に対して，後者が -5.25 ± 0.96 であった．当該組織の著者による論文の確信度はそうでないものよりも高い値で分布しており，識別を実施する際のひとつのツールとして利用できる可能性がある．理想的には，当該組織の著者による論文の確信度が大きくなり，そうでないものが小さくなって完全に分離できればよいが，そのようにはなっていない．本研究においては，データベース内の著者情報について所属組織以外についての属性情報は一切利用しておらず，最も条件の悪い問題設定となっている．当初に述べた通り，この手法と他の手法の組み合わせによって識別性能を高めていくことが必要になる．当該組織の著者でない論文について，当該組織の著者であると判定されているものについては，同姓同名である上に研究分野が類似している，もしくは，研究分野が異なるが論文中に用いられる用語が偶然一致している場合のいずれかが考えられる．前者の場合は，本研究で提案する手法の枠内で改善することは困難であるが，後者は単語を除去する手続きにおいて，逆文書頻度の閾値を最適化することで改善する可能性がある．当該組織の著者の論文について，当該組織の著者と判定されなかった論文については，その著者の通常の分野でない論文や，異分野融合の研究を実施した成果を記述した論文である可能性がある．これらについては今後の課題とする．

図 7 は，同じ条件で，確信度の閾値を変化させた場合の F 値，適合率，再現率をプロットしたものである．前述の閾値で識別した場合には，確信度の上位 583 番目までが当該組織の

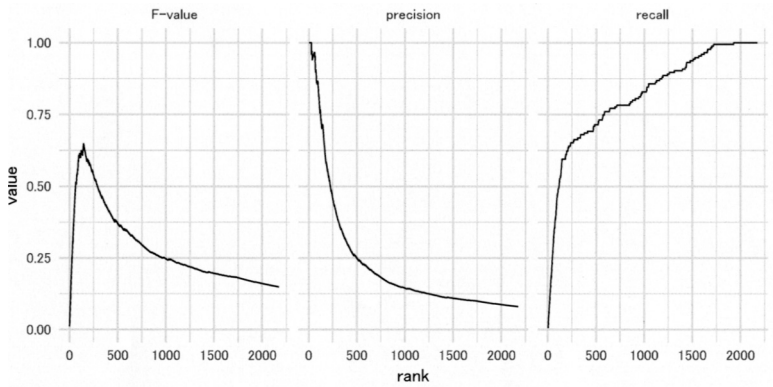


図 7. トピック数を 54 とした CTM による F 値，適合率，再現率.

論文と判定されることになる．この場合の F 値，適合率，再現率はそれぞれ，34.6%，22.5%，74.9% であった．当該組織の著者である論文に対して，確信度が低いものについての改善がなされれば，当該組織の著者でない論文の多くをフィルタリングするツールとしては使えるようになる可能性はある．これについては，トピックモデリングの手法の改善をさらに進める必要がある．また，本研究で比較した LDA，CTM，DMR を統合したような手法として，Structual Topic Modeling (Roberts et al., 2013)が提案されており，これを利用することによって改善する可能性もある．

謝　辞

本研究は平成 29 年度統計数理研究所共同利用研究重点テーマ 2「学術文献 DB における著者識別問題と研究組織評価への応用に関する研究」(29-共研-4201)および平成 30 年度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 のデータベースはこの重点テーマの下で利用許可を受けている．クラリベイト・アナリティクス社をはじめ，データ利用のために尽力していただいた関係者の皆様に謝意を表する．

参 考 文 献

Blei, D. M. (2012). Probabilistic topic models, *Communications of ACM*, **55**(4), 77–84.

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).

Lafferty, J. D. and Blei, D. M. (2006). Correlated topic models, *Advances in Neural Information Processing Systems*, 147–154, MIT Press, Cambridge, Massachusetts.

Lin, C. and He, Y. (2009). Joint sentiment topic model for sentiment analysis, *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 375–384, ACM, New York.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit, http://mallet.cs.umass.edu.

Mimno, D. and McCallum, A. (2008). Topic models conditioned on arbitrary features with dirichlet-multinomial regression, *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, 411–418, AUAI Press, Arlington, Virginia.

Neo4j, Inc (2019). Neo4j Database, https://neo4j.com/neo4j-graph-database/.

Roberts, M. E., Stewart, B. M., Tingley, D., Airoldi, E. M., et al. (2013). The structural topic model and applied social science, *Advances in Neural Information Processing Systems Workshop on Topic Models*: *Computation, Application, and Evaluation*, 1–20, Harrahs and Harveys, Lake Tahoe, Nevada.

Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P. (2004). The author-topic model for authors and documents, *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 487–494, AUAI Press, Arlington, Virginia.

佐藤一誠 (2015).『トピックモデルによる統計的潜在意味解析』, コロナ社, 東京.

Silge, J. and Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in R, *The Journal of Open Source Software*, **1**(3), DOI: http://dx.doi.org/10.21105/joss.00037.

Tang, L. and Walsh, J. (2010). Bibliometric fingerprints: Name disambiguation based on approximate structure equivalence of cognitive maps, *Scientometrics*, **84**(3), 763–784.

218 Proceedings of the Institute of Statistical Mathematics Vol. 68, No. 2, 209–218 (2020)

Author Identification for Scientific Database with Topic Modeling and Its Performance Comparison

Tomokazu Fujino[1] and Hiroka Hamada[2]

[1]International College of Arts and Sciences, Fukuoka Women's University
[2]The Institute of Statistical Mathematics

We propose a method for extracting a list of articles from a scientific literature database whose authors are researchers at a specific organization. The method uses topic modeling, a technique for statistical natural language processing. Topic modeling is applied to papers in which the organization name is included, and feature vectors for each author are created. Based on this, author identification is performed, including the names of researchers in the organization and not containing the organization name. We compared this discrimination performance between several topic models such as Latent Dirichlet Allocation, Dirichlet Multinomial Regression, and Correlated Topic Model.

Key words: Scientific database, evaluating research performance, statistical natural language processing, institutional research.

統計数理(2020)
第 68 巻 第 2 号 219–231

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

トピックモデルを用いた研究動向の分析

武井 美緒[1]・藤野 友和[2]・中野 純司[3,4]

(受付 2019 年 5 月 30 日;改訂 2020 年 1 月 6 日;採択 1 月 16 日)

要　　旨

少子高齢化等に伴い大学の経営難が問題になっている．そのため，大学においても戦略的に学内の支援対象を選択する必要に迫られており，その際には，研究活動の状況や特徴を把握し，評価しなければならない．研究評価の手法としてインパクト・ファクター等の論文の引用情報を用いた手法が利用されることが多い．しかし，引用分析にはいくつかの問題が指摘されている．そこで，研究内容が直接的に表現されている論文の要旨を用い，Hierarchical Dirichlet Process(HDP)を Latent Dirichlet Allocation(LDA)に適用したモデルを利用して要旨内のトピックを抽出し，対象とする組織やグループ毎の研究の動向を把握するための分析方法を紹介する．本研究では，統計科学分野の著名な論文誌と統計科学に関連する二つの研究所の論文の要旨を用いて分析を行い，研究の特徴や論文の発行年度毎の動向が把握できることを確認した．

キーワード：トピックモデル，ノンパラメトリックベイズ，階層ディリクレ過程，Institutional Research.

1. はじめに

国立大学法人運営交付金の削減や少子高齢化に伴い，大学の経営難が問題になっている．そのため，大学においても効果的また，戦略的に学内の支援対象を選択する必要に迫られている．その際には，大学内外や特定の分野の研究活動の状況や特徴を把握し，評価しなければならない．

研究評価の手法としてインパクト・ファクター等の論文の引用情報を用いた手法が利用されることが多い．しかし，引用分析にはいくつかの問題が指摘されている（Cole and Cole, 1971; Porter, 1977; Edge, 1979; Lindsey, 1989）．そこで，研究内容が直接的に表現されている論文の本文(または要旨)を用いた分析が行われており，語の出現頻度や共語(語の共起回数の頻度)を利用し，論文間の研究内容の遠近やクラスター分け等の分析が行われている（Callon et al., 1983; Law et al., 1988; Braam et al., 1991）．

また，論文の要旨を用いた研究として，Proceedings of the National Academy of Science (PNAS)の論文の要旨に対して，トピックモデルを利用し推定したトピックと既存の分類の比較や，トピック毎の流行の調査を行なっているものがある（Griffiths and Steyvers, 2004）．オペ

[1] 統計数理研究所 特任技術専門員：〒 190–8562 東京都立川市緑町 10–3
[2] 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
[3] 中央大学 国際経営学部：〒 192–0393 東京都八王子市東中野 742–1
[4] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3

レーションリサーチや経営科学，交通研究の分野では，対象の分野の論文誌の論文の要旨を用いて，トピックモデルを利用し，要旨の内容をいくつかのトピックに分類し，分野内でどのようなトピックの研究が行われているか，さらにその動向を調査する分析が行われている（Gatti et al., 2015; Sun and Yin, 2017）．これらの研究では，トピック推定にトピックモデルの手法として最もよく利用されている Latent Dirichlet Allocation（LDA）（Blei et al., 2003）を利用している．また，Science の論文を用い，時系列の情報を加味したトピックモデルである Dynamic Topic Model を評価し，トピックの時間変化を調査した研究（Blei and Lafferty, 2006）や，時系列の情報とノンパラメトリックベイズの手法の 1 つである，Hierarchical Dirichlet Process（HDP）（Teh et al., 2006）を拡張したモデルを提案し，Neural Information Processing Systems（NIPS）の論文の時系列の特徴を調査した研究がある（Ahmed and Xing, 2010）．

本研究では論文の要旨を用いて，研究評価のために必要な対象の組織やグループ毎の研究の動向を把握することを目的とする．対象の組織やグループの全ての要旨を HDP を利用していくつかのトピックに分類し，グループ毎に集計し，トピックの動向や特徴を探る．

2. モデリング手法

利用するモデル及びそのサンプリング手法について説明する．

2.1 Hierarchical Dirichlet Process（HDP）

HDP はノンパラメトリックベイズの手法の 1 つで，Dirichlet Process（DP）を階層化し，子の DP が親の DP から得られた分布を基底分布とすることで，各状態で情報を共有することができる．そのため，HDP を LDA に適用することで文書間でトピックを共有することが可能になる．今後，HDP を LDA に適用した場合を HDP-LDA と呼ぶ．HDP-LDA の生成過程は以下の様になる（Teh and Jordan, 2010）．

(1) $G_0 \mid \gamma, H \sim DP(\gamma, H)$
(2) For each document $d = 1, \ldots, D$
  (a) $G_d \mid \alpha, G_0 \sim DP(\alpha, G_0)$
  (b) For each word $n = 1, \ldots, N_d$
    i. $\theta_{dn} \mid G_d \sim G_d$
    ii. $w_{dn} \mid \theta_{dn} \sim F(\theta_{dn})$

ここで，H はパラメータ β の Dirichlet 分布，w_{dn}，θ_{dn} はそれぞれ文書 d の n 番目の単語，トピック．α，γ はハイパーパラメータを表す．$F(\theta_{dn})$ は単語毎の分布で多項分布を取る．

文書生成モデルを評価する指標として Perplexity がよく利用され，以下の式で定義される（Teh et al., 2006）．

$$\exp\left(-\frac{1}{I}\log p(w_1, \ldots, w_I|\text{Training corpus})\right) \tag{2.1}$$

ここで，$p(\cdot)$ は対象のモデルの確率関数，I はテストデータの単語数を表す．

2.2 Chinese Restaurant Franchise（CRF）

HDP-LDA のサンプリングには理解が容易な Chinese restaurant process（CRP）を拡張した Chinese restaurant franchise（CRF）（Teh et al., 2006）を使用した．

CRF は CRP を共通の料理をもつ複数のレストランへ拡張したサンプリング手法である．CRF ではメニューが共有されているレストランのフランチャイズを考える．それぞれのレス

表 1. 記号の説明.

記号	説明
D	文書（レストラン）数
K	トピック（料理）数
M	テーブル数
M_k	トピック（料理）k を選んだテーブル数
N	全文書（レストラン）内の単語（客）数
N_d	文書（レストラン）d 内の単語（客）数
N_k	文書（レストラン）全体でトピック（料理）k が選ばれたテーブルに属する単語（客）数
N_{dk}	文書（レストラン）d でトピック（料理）k が選ばれたテーブルに属する単語（客）数
N_{dl}	文書（レストラン）d のテーブル l を選んだ単語（客）数
N_{kv}	文書（レストラン）全体でトピック（料理）k が選ばれたテーブルに属する語彙 v の数
N_{dlv}	文書（レストラン）d のテーブル l に属する語彙 v の数
T	単語（客）が属するテーブル集合
t_{dn}	文書（レストラン）d の n 番目の単語（客）のテーブル
V	文書（レストラン）全体の語彙数
W	文書（レストラン）集合
w_{dn}	文書（レストラン）d の n 番目の単語（客）
Z	選ばれたトピック（料理）集合
z_{dl}	文書（レストラン）d のテーブル l にて選ばれたトピック（料理）
z_{dn}	文書（レストラン）d の n 番目の単語（客）の選んだトピック（料理）
θ_{dk}	文書（レストラン）d でトピック（料理）k が選ばれる確率
ϕ_{kv}	語彙 v がトピック（料理）k を選ぶ確率
α	文書（レストラン）毎の DP のハイパーパラメータ
γ	文書（レストラン）集合全体の DP のハイパーパラメータ
β	トピック（料理）毎の単語（客）分布のハイパーパラメータ

トランに無限個のテーブルがあり，客がレストランに入店した際にテーブルを選び，選んだテーブルに客がいなければ料理を 1 つ選ぶ．他の客がいる場合は事前に選ばれている料理が提供される．どのテーブルにつくかの確率はそのテーブルを選んだ客の数に比例し，どの料理を選ぶかの確率はその料理を選んでいるレストラン全体のテーブルの数に比例する．なお，テーブルにて選択される料理は 1 つ，複数のテーブルで同じ料理を選択することが可能である．トピックモデルとの対応として，レストランが文書，テーブルにて選択された料理がトピック（トピック数はレストラン全体の料理の種類数），客が各文書の単語となる．

CRF による HDP-LDA のサンプリングはモデル内のパラメータを周辺化削除した上でサンプリングを行う周辺化ギブスサンプリングを用いる．本研究では，事前分布としてパラメータ β の対称 Dirichlet 分布を用いた．表記法を表 1 に，推定方法を以下に示す（岩田, 2015）．

文書（レストラン）d の n 番目の単語（客）がテーブル l を選ぶ確率は，以下となる．

$$
\begin{aligned}
(2.2)\quad p(t_{dn}=l \mid W, T_{\backslash dn}, Z, \alpha, \gamma, \beta) &= \begin{cases} p(t_{dn}=l, z_{dl}=k \mid W, T_{\backslash dn}, Z, \alpha, \gamma, \beta) \\ p(t_{dn}=l_{new}, z_{dl_{new}}=k \mid W, T_{\backslash dn}, Z, \alpha, \gamma, \beta) \\ p(t_{dn}=l_{new}, z_{dl_{new}}=k_{new} \mid W, T_{\backslash dn}, Z, \alpha, \gamma, \beta) \end{cases} \\
&\propto \begin{cases} N_{dl\backslash dn} \dfrac{N_{z_{dl} w_{dn} \backslash dn} + \beta}{N_{z_{dl}\backslash dn} + \beta V} & \text{既存テーブル} \\ \alpha \dfrac{M_k}{M+\gamma} \dfrac{N_{k w_{dn}\backslash dn} + \beta}{N_{k\backslash dn} + \beta V} & \text{新テーブル，既存トピック } k \\ \alpha \dfrac{\gamma}{M+\gamma} \dfrac{1}{V} & \text{新テーブル，新トピック} \end{cases}
\end{aligned}
$$

文書(レストラン)d の l 番目のテーブルにてトピック(料理)k が選ばれる確率は，以下となる．

$$(2.3)\quad p(z_{dl}{=}k|W,T,Z_{\backslash dl},\gamma,\beta){=}\begin{cases}p(z_{dl}{=}k|W,T,Z_{\backslash dl},\gamma,\beta)\\ p(z_{dl}{=}k_{new}|W,T,Z_{\backslash dl},\gamma,\beta)\end{cases}$$
$$\propto\begin{cases}M_{k\backslash dl}\dfrac{\Gamma(N_{k\backslash dl}+\beta V)}{\Gamma(N_{k\backslash dl}+N_{dl}+\beta V)}\prod_{v=1}^{V}\dfrac{\Gamma(N_{kv\backslash dl}+N_{dlv}+\beta)}{\Gamma(N_{kv\backslash dl}+\beta)} & \text{既存トピック}\\ \gamma\dfrac{\Gamma(\beta V)}{\Gamma(N_{dl}+\beta V)}\dfrac{\prod_{v=1}^{V}\Gamma(N_{dlv}+\beta)}{\Gamma(\beta)^{V}} & \text{新トピック}\end{cases}$$

ここで，$T_{\backslash dn}$ や $N_{dl\backslash dn}$ などの下付き文字でバックスラッシュが付いている集合や数は，下付き文字で示されている値を除いた集合や数を表す．例えば，$T_{\backslash dn}=\{t_{11},\ldots,t_{d,n-1},t_{d,n+1},\ldots,t_{DN_d}\}$ は T から $t_{\backslash dn}$ を除いた集合を表す．また，l_{new}，k_{new} は新しく選ばれたテーブル，料理を表す．(2.2)と(2.3)を繰り返しサンプリングを行う．

反復後，文書トピック確率 θ_{dk} とトピック単語確率 ϕ_{kv} の推定値は以下のように計算できる．

$$(2.4)\qquad \theta_{dk}=\begin{cases}\dfrac{1}{N_d+\alpha}\left(N_{dk}+\alpha\dfrac{M_k}{M+\gamma}\right) & \text{既存トピック}\\ \dfrac{1}{N_d+\alpha}\dfrac{\alpha\gamma}{M+\gamma} & \text{新トピック}\end{cases}$$

$$(2.5)\qquad \phi_{kv}=\begin{cases}\dfrac{N_{kv}+\beta}{N_k+\beta V} & \text{既存トピック}\\ \dfrac{1}{V} & \text{新トピック}\end{cases}$$

3. 分析方法

本研究の分析方法を説明する．分析方法の概要を図 1 に示す．

3.1 データ前処理

対象のグループ全ての論文の要旨を単語に分割し，データの前処理を行い，単語頻度表を作成する．

3.2 モデリング

単語頻度表のデータを元に，HDP-LDA を利用しパラメータ推定を行い，要旨の内容をデー

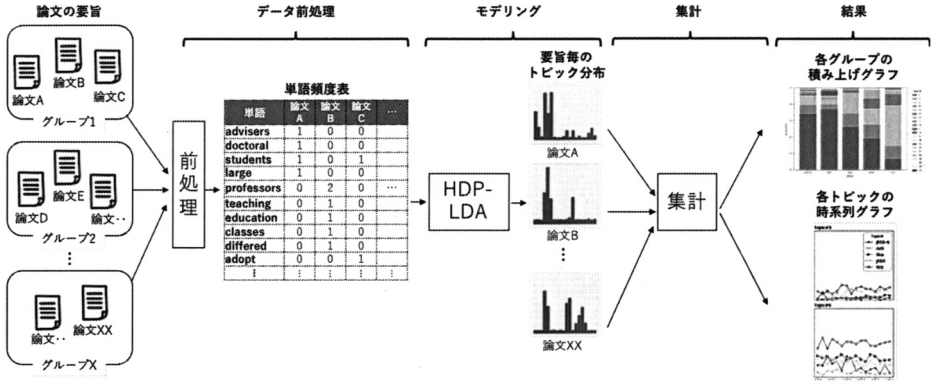


図 1．分析方法の概要．

タから推定した特定の数のトピックに分類する．推定後，文書トピック確率 θ_{dk} とトピック単語確率 ϕ_{kv} を計算する．

3.3 集計

結果として得られた文書トピック確率 θ_{dk} と論文 d の発行年度，グループの情報を用いて，発行年度及びグループ毎に集計を行う．

発行年度 t，グループ g，トピック k のトピック割合 p_{tgk} は，

$$p_{tgk} = \frac{\sum_{d=1}^{D} \theta_{dk} \times \delta(t_d = t \wedge g_d = g)}{D_{tg}} \tag{3.1}$$

と計算する（Gatti et al., 2015; Sun and Yin, 2017）．ここで，t_d は文書 d の発行年度，g_d は文書 d のグループ，D_{tg} は発行年度 t，グループ g の文書数を表す．また，$\delta(x)$ は x が正の場合 1 を，それ以外の場合は 0 を返す関数である．

4. 分析

本研究では，統計科学の分野の著名な論文誌と統計科学に関連する研究所の論文の要旨を用いて分析を行った．分析用の論文の要旨，発行年度の情報は Web of Science から取得した．HDP-LDA の実装は Nakatani Shuyo 氏のコードを参考にした（https://github.com/shuyo/iir/blob/master/lda/lda.py）．データの前処理として，単語の小文字化，活用形の統一のために Lemmatization（単語を辞書の見出し語のような活用されていない形に変換すること）の実施，ストップワード，記号，数字，1 文字の単語の除去を行った．また，全ての要旨の中で一度しか出現しない単語は分析に影響を与えないと考え取り除いた．さらに，極端に出現回数が多い単語は，分析用データの要旨を記載する際に共通の単語であり，要旨の内容を判別する際には不要と考え，それぞれのデータ毎に単語の出現回数を確認した上で，特定の回数以上出現した単語は取り除いた．パラメータ推定の際に，ハイパーパラメータは，α，γ は分布 $Gamma(1,1)$ に従うとし，$\beta = 0.5$ とした．反復は反復回数毎の対数尤度の時系列プロットを用いて対数尤度の値が安定するまで繰り返した（Omori, 2001）．対数尤度が安定した後の試行の中から対数尤度が最大の時のパラメータを利用した．本手法の検証として，各データセットの単語の 90% を訓練データ，残りの 10% をテストデータとして分割し，HDP-LDA と LDA でそれぞれ 3 回ずつ分析を行い，perplexity の値を比較し，HDP-LDA での分析結果が LDA の結果と比べて予測性能が悪化していないことを確認した．結果を表 2 に示す．LDA のトピック数は各 HDP-LDA の試行で推定された値，ハイパーパラメータ α と β は HDP-LDA と同一の対称 Dirichlet 分布を用いた．

4.1 統計科学の分野の著名な論文誌の分析

統計科学の著名な論文誌 5 誌（Varin et al., 2016）について，発行年度が 2001 年から 2016 年までの要旨を用い，各論文誌の研究の特徴や動向を把握するための分析を行った．論文誌の名称と論文数を表 3 に示す．

表 2. HDP-LDA と LDA の Perplexity の比較.

	Perplexity 平均値	
	HDP-LDA	LDA
統計科学論文誌データ	1,764.68	1,826.03
研究所データ	1,473.72	1,644.70

表 3. 統計科学の著名な論文誌の名称と論文数.

雑誌名	略称	論文数
JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY	JRSS-B	657
ANNALS OF STATISTICS	AoS	1,573
BIOMETRIKA	Bka	1,217
JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION	JASA	1,873
BIOMETRICS	Bcs	2,054
	合計	7,374

データの前処理として，本データでは 4,000 回以上出現した単語は取り除いた(取り除いた単語は次の通り．model, data, method, use, propose, study, estimator, estimate, test, distribution, sample)．単語の種類数は V=11,114，単語の総数は N=545,091 となった．HDP-LDA を用いたパラメータ推定を反復回数 3,000 回行い，その結果トピック数は 15 となった．

推定されたパラメータを用いて計算したトピック割合 p_{tgk} をトピック毎に集計し，集計した値が上位のトピックのトピック単語確率 ϕ_{kv} を表 4 に，そのトピックの時系列グラフを図 2 に示す．また，2016 年度の各論文誌のトピック割合 p_{tgk} の積み上げグラフを図 3 に示す．

表 4 及び文書トピック確率 θ_{dk} の値が高かった要旨のタイトルより，トピック 1 は変数選択やモデル選択に関連した分析手法についてのトピック，トピック 2 は密度推定等の推定手法やその漸近特性に関連するトピック，トピック 3 は遺伝子を用いた分析に関連するトピック，トピック 4 は空間モデリング等の統計モデルを利用した分析に関連したトピック，トピック 5 は臨床試験の際に用いられる用量設定試験や因果推論に関連するトピック，トピック 6 は疫学研究や機械の故障等の推定で用いられる生存時間解析に関連するトピック，トピック 8 は仮説検定のエラー率の調整に関連した分析手法についてのトピックと考えられる．また，図 3 より，BIOMETRICS は他の論文誌に比べてトピック 3，5，6 の遺伝子や疫学研究に関連するトピックの割合が大きいことがわかる．BIOMETRICS は統計学や数学のバイオサイエンスへの応用に関連する記事を掲載しているジャーナル (Biometrics mission, 2019)のため現実に即した結果であることが確認できる．図 2 より，変数選択等に関連したトピック 1 は全ての論文誌でトピック割合が増加傾向であることがわかる．近年のデータの飛躍的増大に伴い，変数選択に関連した研究が盛んに行われている結果であると考えられる．また，トピック 2 のグラフでは，他の論文誌に比べ ANNALS OF STATISTICS のトピック割合が例年高くなっていることがわかる．推定に関連した内容の論文が対象の論文誌ではよく取り上げられている可能性があると考えられる．本分析を用いることで，各論文誌の特徴，各トピックの動向から論文誌の全体の研究の動向，各論文誌の掲載論文のトピックの違いが把握できることを確認した．

4.2 研究所の分析

統計数理研究所(以下 ISM)と Academia Sinica 統計科学研究所(以下 Academia Sinica) の論文の要旨を用い，二つの研究所の研究の特徴や動向を把握するための分析を行った．ISM と Academia Sinica のデータとして，著者にそれぞれの機関に所属している人が 1 人以上含まれる論文を対象の機関のデータとした．また，発行年度は 2001 年から 2016 年までのデータを利用した．各研究所の論文数を表 5 に示す．

データの前処理として，本データでは 1,000 回以上出現した単語は取り除いた(取り除いた単語は次の通り．model, method, use, data, study)．単語の種類数は V=7,288，単語の総数は N=120,704 となった．HDP-LDA を用いたパラメータ推定を反復回数 1,000 回行い，トピック数は 18 となった．

推定されたパラメータを用いて計算したトピック割合 p_{tgk} をトピック毎に集計し，集計した

表 4. 統計科学論文誌データのトピック割合上位 7 トピックの単語確率 ϕ_{kv} の上位 20 単語.

Topic1		Topic2		Topic3		Topic4		Topic5		Topic6		Topic8	
Word	ϕ_{kv}	Word	ϕ_{kv}	Word	ϕ_{kv}	Word	ϕ_{kv}	Word	ϕ_{kv}	Word	ϕ_{kv}	Word	ϕ_{kv}
regression	0.01360	process	0.01300	gene	0.02049	spatial	0.01223	design	0.03476	time	0.01603	statistic	0.01859
parameter	0.01090	function	0.01284	genetic	0.00960	process	0.01054	treatment	0.03146	effect	0.01021	procedure	0.01622
function	0.01074	matrix	0.01047	expression	0.00890	population	0.00853	effect	0.02271	survival	0.00935	hypothesis	0.01360
variable	0.01053	density	0.01028	analysis	0.00793	time	0.00846	trial	0.01638	approach	0.00897	confidence	0.01036
approach	0.00981	result	0.00979	identify	0.00670	approach	0.00658	outcome	0.01123	covariates	0.00872	error	0.01023
estimation	0.00936	asymptotic	0.00921	multiple	0.00657	bayesian	0.00604	clinical	0.00868	event	0.00870	interval	0.00998
linear	0.00872	rate	0.00901	number	0.00639	article	0.00552	patient	0.00846	regression	0.00833	size	0.00921
algorithm	0.00836	time	0.00831	approach	0.00628	survey	0.00511	causal	0.00777	simulation	0.00759	null	0.00903
simulation	0.00808	covariance	0.00716	image	0.00626	state	0.00481	randomize	0.00655	risk	0.00737	power	0.00902
selection	0.00788	estimation	0.00669	statistical	0.00566	markov	0.00481	dose	0.00645	estimation	0.00733	base	0.00842
problem	0.00782	paper	0.00658	sequence	0.00545	develop	0.00476	optimal	0.00636	analysis	0.00724	bootstrap	0.00774
result	0.00728	condition	0.00648	association	0.00503	statistical	0.00463	group	0.00629	longitudinal	0.00722	rate	0.00765
procedure	0.00698	series	0.00644	develop	0.00490	individual	0.00443	response	0.00499	hazard	0.00705	control	0.00734
likelihood	0.00675	problem	0.00633	trait	0.00482	rate	0.00443	control	0.00462	censor	0.00700	ratio	0.00727
prior	0.00634	class	0.00618	apply	0.00451	change	0.00433	analysis	0.00430	disease	0.00695	variance	0.00712
analysis	0.00621	case	0.00607	microarray	0.00445	information	0.00425	level	0.00394	miss	0.00657	likelihood	0.00687
new	0.00610	consider	0.00593	cell	0.00440	effect	0.00420	subject	0.00389	covariate	0.00652	result	0.00665
base	0.00606	gaussian	0.00593	article	0.00440	analysis	0.00400	compare	0.00385	outcome	0.00600	simulation	0.00662
component	0.00605	convergence	0.00585	structure	0.00435	count	0.00396	result	0.00381	function	0.00599	alternative	0.00605
property	0.00578	random	0.00542	cluster	0.00432	probability	0.00387	experiment	0.00379	measurement	0.00583	asymptotic	0.00596

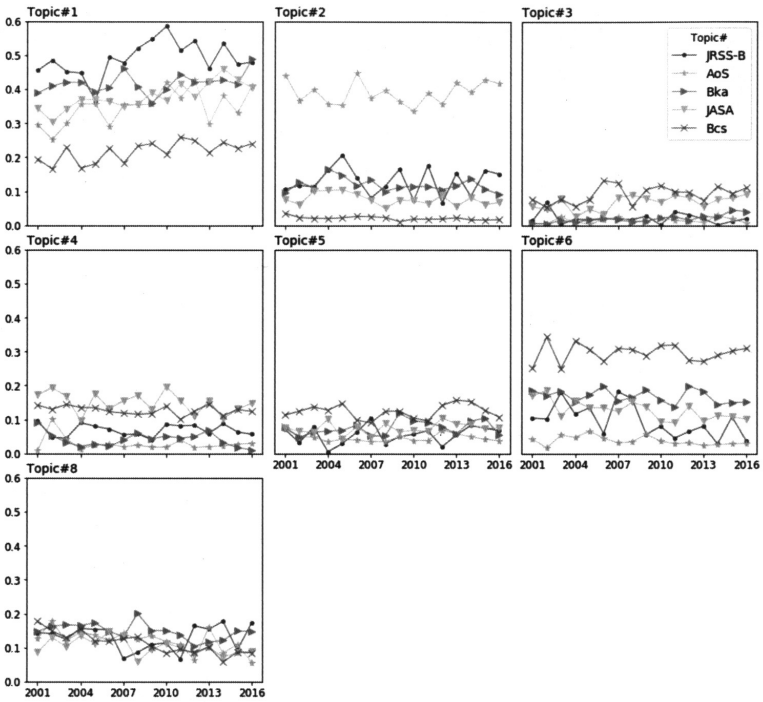


図 2. 統計科学論文誌データのトピック割合の時系列グラフ.

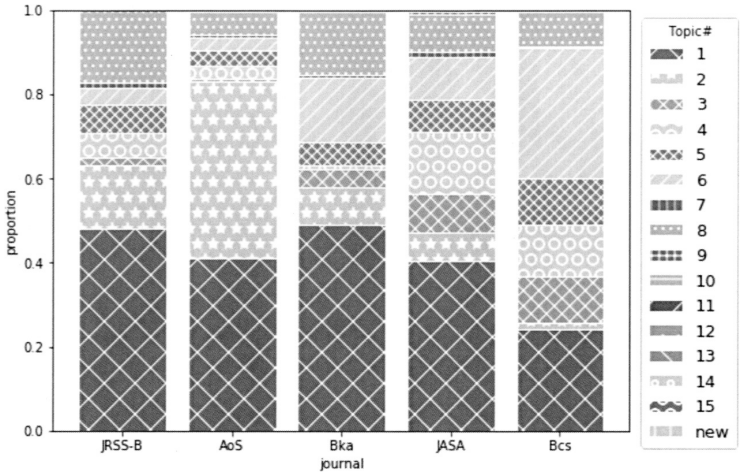


図 3. 統計科学論文誌データの 2016 年度のトピック割合の積み上げグラフ.

値が上位のトピックのトピック単語確率 ϕ_{kv} を表 6 に，そのトピックの時系列のグラフを図 4 に示す．また，2016 年度の各研究所のトピック割合 p_{tgk} の積み上げグラフを図 5 に示す．

表 6 及び文書トピック確率 θ_{dk} の値が高かった要旨のタイトルより，トピック 1 は統計科学の論文に共通して登場する単語が集まったトピック，トピック 2 はミトコンドリア等を用いた種の起源や系統等の分析に関連するトピック，トピック 3 は生物医学分野の細胞や遺伝子を用

表 5. 各研究所の論文数.

研究所名	論文数
ISM	1,003
Academia Sinica	595
合計	1,598

表 6. 研究所データのトピック割合上位 6 のトピック単語確率 ϕ_{kv} の上位 20 単語.

Topic1		Topic2		Topic3		Topic4		Topic6		Topic7	
Word	ϕ_{kv}	Word	ϕ_{kv}	Word	ϕ_{kv}	Word	ϕ_{kv}	Word	ϕ_{kv}	Word	ϕ_{kv}
propose	0.01448	specie	0.01571	cell	0.02394	patient	0.00950	earthquake	0.02294	image	0.01459
distribution	0.01133	population	0.01300	gene	0.02327	risk	0.00835	aftershock	0.01143	brain	0.00650
result	0.00780	sequence	0.01142	cancer	0.01972	associate	0.00654	event	0.00894	eeg	0.00557
test	0.00742	gene	0.01114	expression	0.01658	health	0.00567	region	0.00833	task	0.00463
approach	0.00734	tree	0.00753	lung	0.01035	disease	0.00560	forecast	0.00809	subject	0.00463
base	0.00734	phylogenetic	0.00663	patient	0.00713	factor	0.00560	rate	0.00712	fmri	0.00451
parameter	0.00726	genome	0.00635	identify	0.00579	association	0.00546	time	0.00676	neuron	0.00381
problem	0.00726	genetic	0.00522	protein	0.00553	effect	0.00495	seismicity	0.00670	sensor	0.00357
paper	0.00724	suggest	0.00494	mutation	0.00486	treatment	0.00466	stress	0.00664	dynamic	0.00334
function	0.00659	relationship	0.00409	tumor	0.00479	group	0.00452	change	0.00627	region	0.00322
algorithm	0.00641	analysis	0.00409	analysis	0.00432	year	0.00437	sequence	0.00609	activity	0.00322
estimate	0.00637	individual	0.00409	target	0.00425	age	0.00430	variation	0.00536	phase	0.00322
process	0.00621	evolution	0.00398	microarray	0.00425	subject	0.00379	japan	0.00506	signal	0.00311
analysis	0.00601	group	0.00398	pathway	0.00412	ci	0.00372	current	0.00488	respiratory	0.00299
time	0.00600	evolutionary	0.00370	effect	0.00378	increase	0.00358	slip	0.00482	network	0.00287
sample	0.00547	region	0.00353	egfr	0.00372	exposure	0.00351	magnitude	0.00470	pixel	0.00287
estimation	0.00524	analyse	0.00341	treatment	0.00372	score	0.00322	eta	0.00464	connectivity	0.00275
apply	0.00490	mitochondrial	0.00330	level	0.00358	result	0.00322	activity	0.00458	evaluate	0.00275
variable	0.00487	base	0.00324	interaction	0.00352	compare	0.00314	seismic	0.00409	time	0.00275
information	0.00483	map	0.00324	human	0.00352	control	0.00314	result	0.00409	functional	0.00264

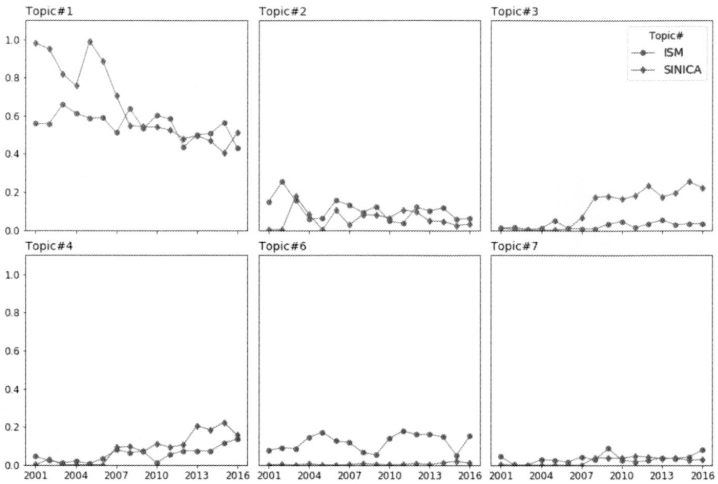


図 4. 研究所データのトピック割合の時系列グラフ.

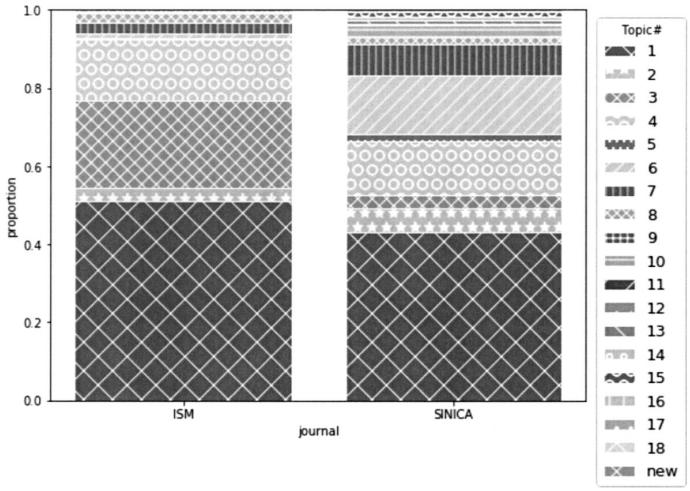


図 5. 研究所データの 2016 年度のトピック割合の積み上げグラフ.

いた分析に関連するトピック，トピック 4 は臨床試験等の経時データの分析に関連するトピック，トピック 6 は地震の分析に関連するトピック，トピック 7 は脳波や MRI を用いた分析に関連するトピックと考えられる．図 5 より，2016 年度では ISM は地震の分析に関連したトピック 6，Academia Sinica は生物医学分野に関連したトピック 3 の割合がお互いと比較すると割合が大きくなっていることがわかる．この結果より，近年の二つの研究所の特徴の違いを確認することができる．さらに，図 4 より，ISM は 2001 年から連続的にトピック 6 の割合が大きいが，Academia Sinica は 2006 年頃からトピック 3 の割合が増加していることがわかる．また，ISM ではトピック 4 の割合が，Academia Sinica ではトピック 3 と 4 の割合が増加傾向にあることから，両研究所で医学統計に関連する研究が増加していると考えられる．本分析では，二つの研究所の特徴，その動向やどのような研究のトピックの割合が増加傾向にあるか把握できることを確認した．

5. おわりに

本研究では，対象の組織やグループの研究の特徴や動向を把握するために，論文の要旨を用い，HDP-LDA を利用したトピック推定の結果を元に分析する手法を紹介した．統計科学の著名な論文誌と研究所のデータを用いて分析を行い，対象のグループの研究の特徴や論文の発行年度毎の動向が把握できることを確認した．

また，今回は研究の動向を把握するため，論文の要旨のデータについての分析を行ったが，本分析手法はテキストデータのような離散型のデータであれば適用できるため，更なる活用が期待できると考えられる．

参 考 文 献

Ahmed, A. and Xing, E. P. (2010). Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream, *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 20–29.

Biometrics mission (2019). http://www.biometrics.tibs.org/, 2019 年 5 月 27 日アクセス.

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

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models, *Proceedings of the 23rd International Conference on Machine Learning*, 113–120, ACM, New York.

Braam, R. R., Moed, H. F. and van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis. I. Structural aspects, *Journal of the American Society for Information Science*, **42**, 233–251.

Callon, M., Courtial, J. P., Turner, W. A. and Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis, *Social Science Information*, **22**, 191–235.

Cole, J. R. and Cole, S. (1971). Measuring the quality of sociological research: Problems in the use of the science citation index, *The American Sociologist*, **6**, 23–29.

Edge, D. (1979). Quantitative measures of communication in science: A critical review, *History of Science*, **17**, 102–134.

Gatti, C. J., Brooks, J. D. and Nurre, S. G. (2015). A historical analysis of the field of OR/MS using topic models, arXiv preprint, arXiv:1510.05154.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics, *Proceedings of the National Academy of Sciences of the United States of America*, **101**(Suppl 1), 5228–5235.

岩田具治 (2015).『機械学習プロフェッショナルシリーズトピックモデル』, 講談社, 東京.

Law, J., Bauin, S., Courtial, J. P. and Whittaker, J. (1988). Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification, *Scientometrics*, **14**, 251–264.

Lindsey, D. (1989). Using citation counts as a measure of quality in science measuring what's measurable rather than what's valid, *Scientometrics*, **15**, 189–203.

Omori, Y. (2001). Recent developments in markov chain monte carlo method, *Journal of the Japan Statistical Society*, **31**(3), 305–344.

Porter, A. L. (1977). Citation analysis: Queries and caveats, *Social Studies of Science*, **7**, 257–267.

Sun, L. and Yin, Y. (2017). Discovering themes and trends in transportation research using topic modeling, *Transportation Research Part C*: *Emerging Technologies*, **77**, 49–66.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006). Hierarchical dirichlet processes, *Journal of the American Statistical Association*, **101**, 1566–1581.

Teh, Y. W. and Jordan, M. I. (2010). Hierarchical bayesian nonparametric models with applications, *Bayesian Nonparametrics* (eds. N. L. Hjort, C. Holmes, P. Müller and S. G. Walker), 158–207, Cambridge University Press, Cambridge.

Varin, C., Cattelan, M. and Firth, D. (2016). Statistical modeling of citation exchange between statistics journals, *Journal of the Royal Statistical Society*: *Series A*, **179**, 1–63.

Understanding Research Trends Based on Article Abstracts Using Topic Modeling

Mio Takei[1], Tomokazu Fujino[2] and Junji Nakano[1,3]

[1]The Institute of Statistical Mathematics
[2]Faculty of International College of Arts and Sciences, Fukuoka Women's University
[3]Faculty of International Economics, Chuo University

The financial difficulties experienced by universities due to declining birth rates and aging populations are becoming a social problem. It is necessary to identify and evaluate the trend of research activities inside and outside universities in order to strategically select support targets in these institutions. Methods in research evaluation often use article citation information such as the impact factor. However, it has been pointed out that there are several problems with this approach. Therefore, we employ a model that applies the Hierarchical Dirichlet Process (HDP) to Latent Dirichlet Allocation (LDA) for the inference of topics using abstracts of articles in which the research content is directly expressed, and show a method for determining the research trend of each target organization and group. We use abstracts from representative journals in the field of statistical sciences and from institutes related to statistical sciences to analyze the method. In the analysis, we confirm that the results can identify the research characteristics for each target group and the research trends for each year of publications.

Key words: Topic modeling, nonparametric Bayesian statistics, Hierarchical Dirichlet Process, institutional research.

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

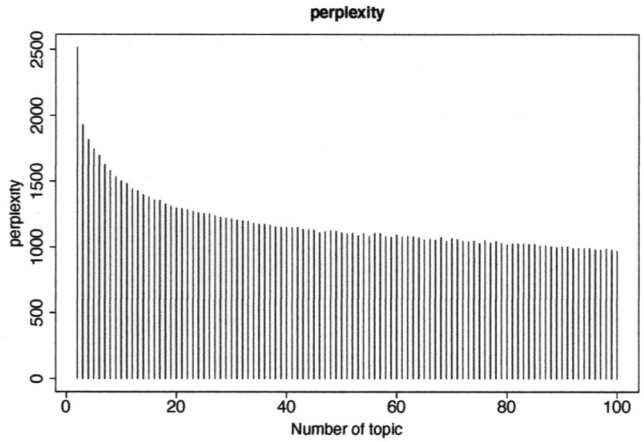


図 1．Perplexity の推移．

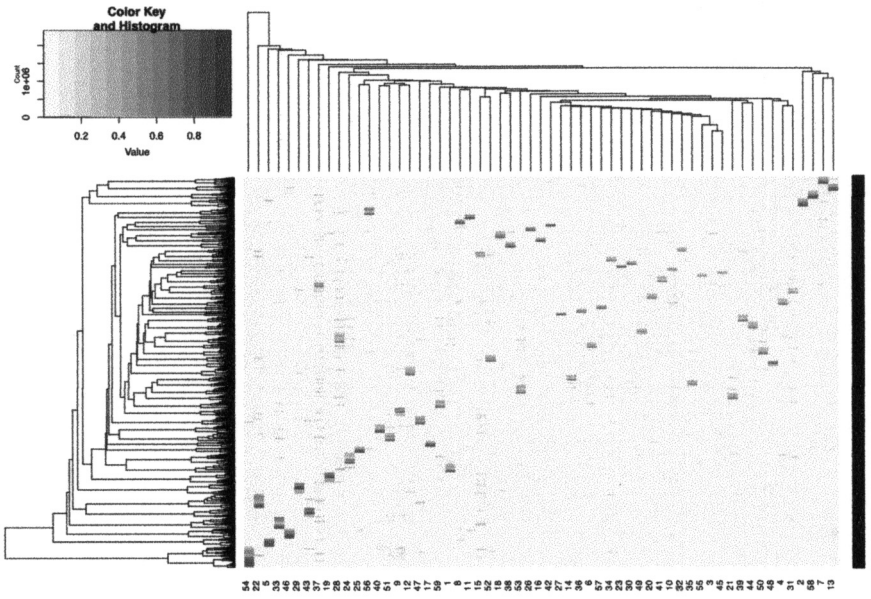


図 2．各論文のトピックの出現確率のヒートマップ．

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

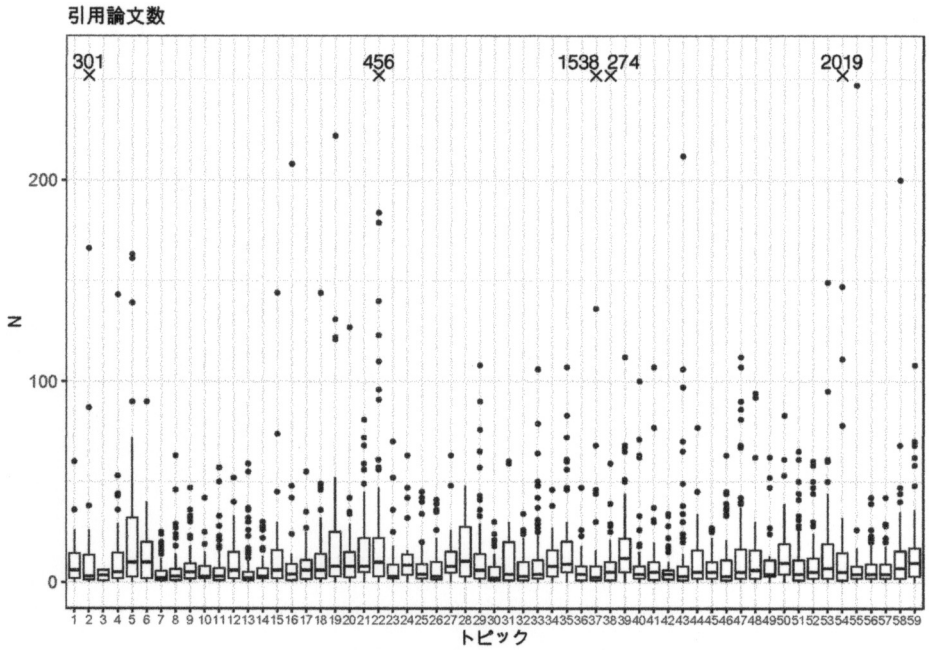


図 3. 引用論文数の箱ひげ図.

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

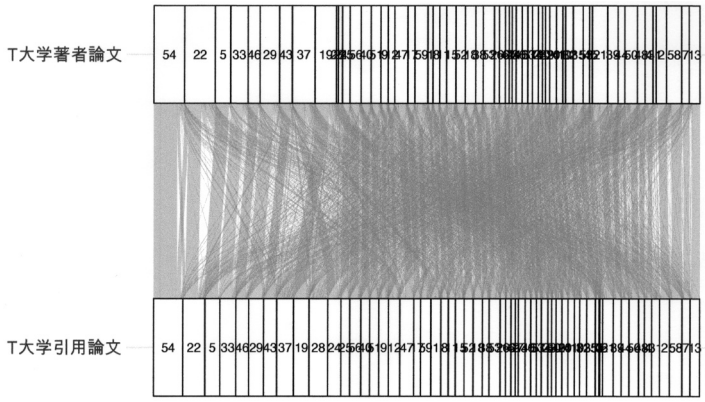


図 4. T 大学著者論文と T 大学引用論文のトピックの関連性.

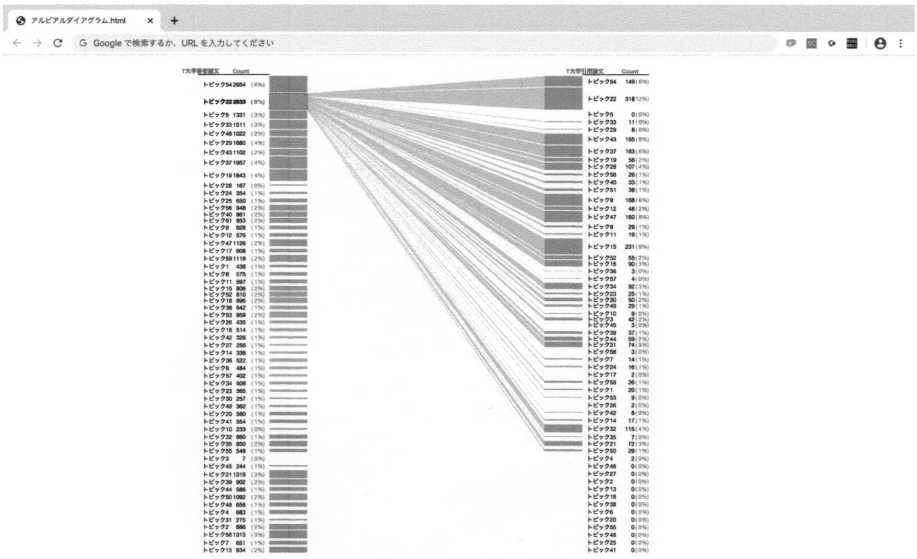


図 5. D3.js による T 大学著者論文と T 大学引用論文のトピックの関連性のインタラクティブな可視化.

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．

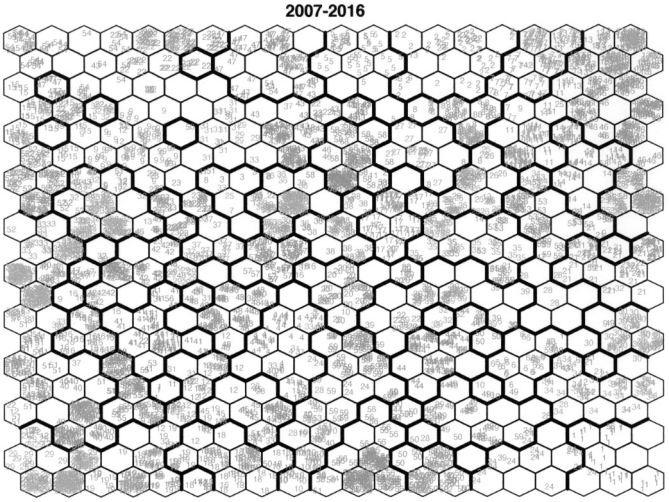


図 6. 入力データを標準化した自己組織化マップによる可視化.

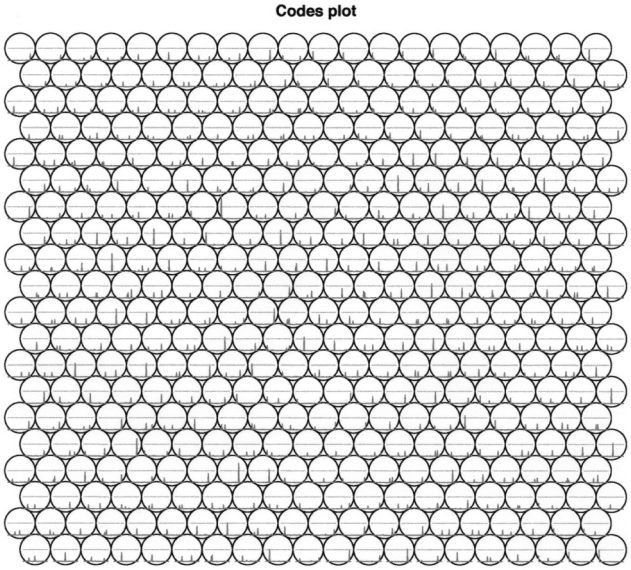


図 7. 自己組織化マップの各ユニットにおける各トピックに対する重みベクトル.

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．

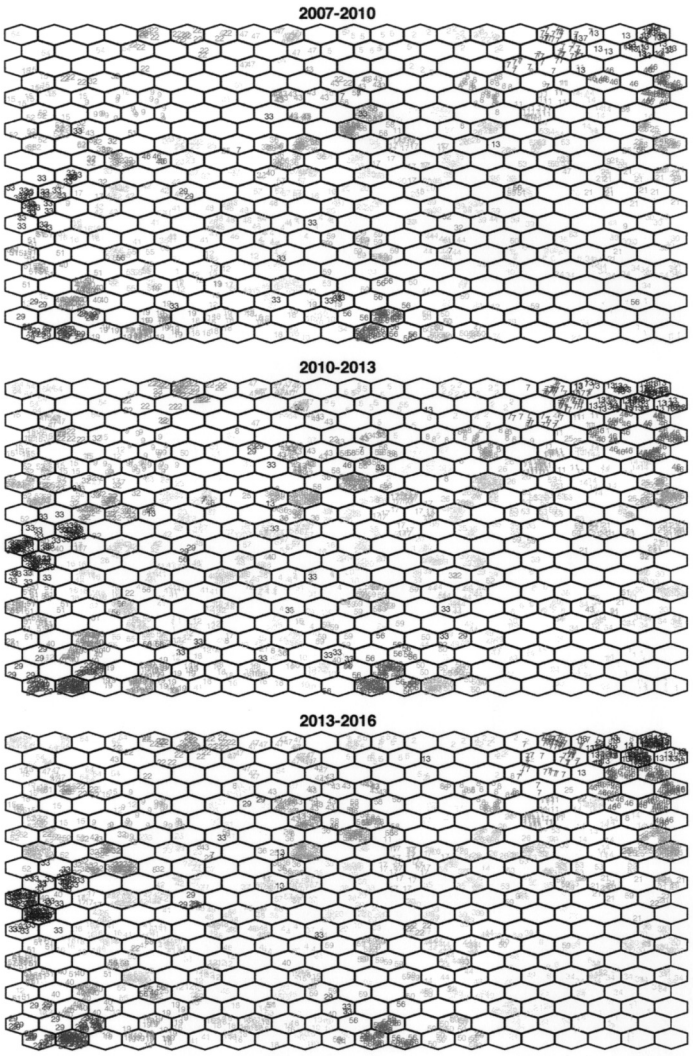


図 8. 経年的な変化の可視化.

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.

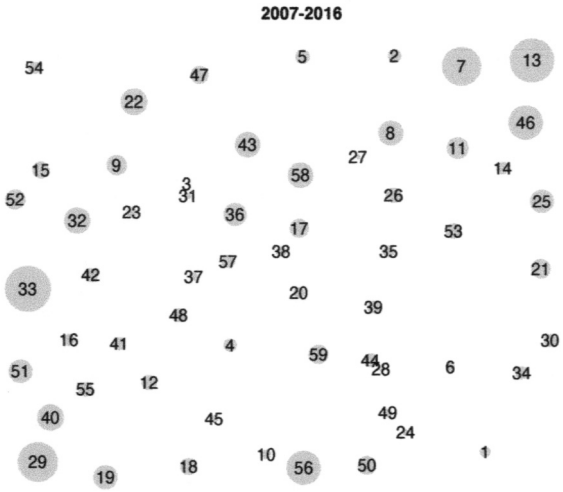


図 9. 各トピックの重心による可視化.

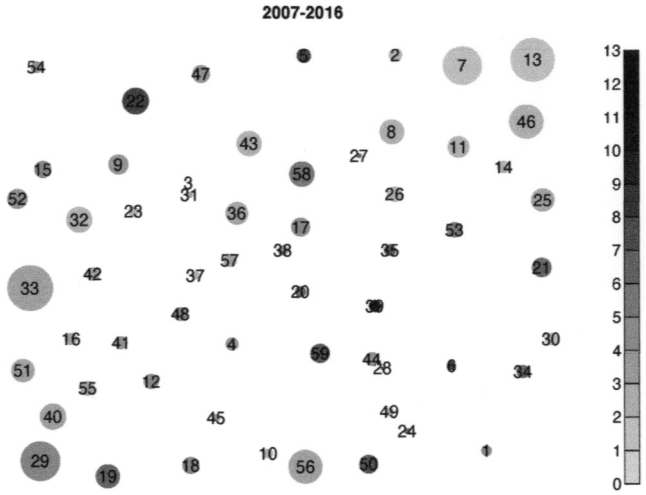


図 10. 被引用論文数の情報を追加した可視化.

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

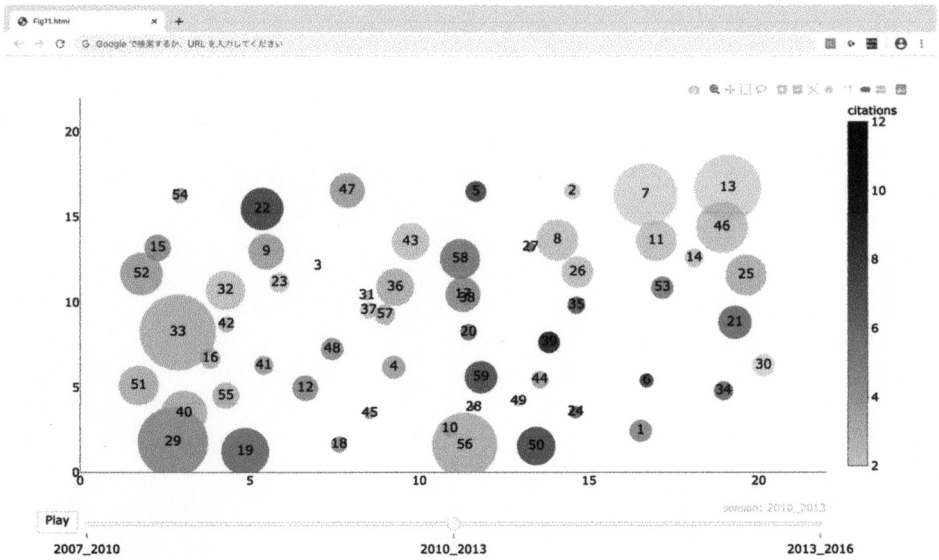


図 11. モーショングラフによる経年変化の可視化.

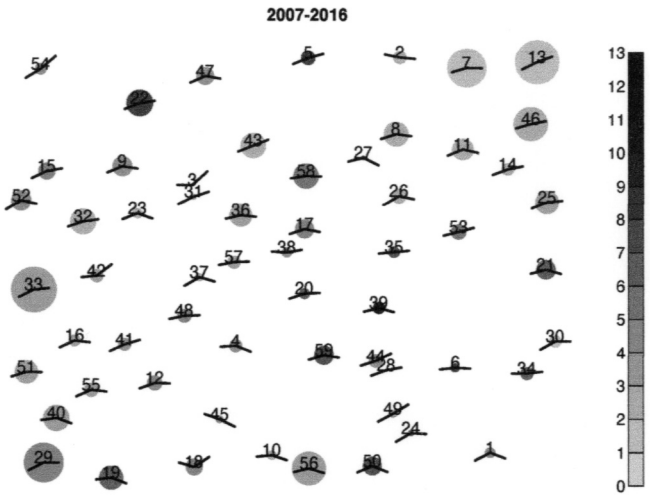


図 12. 経年的な変化の情報を追加した可視化.

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.

統計数理(2020)
第 68 巻 第 2 号 247–264

特集「Institutional Research と統計科学」
［研究ノート］

学術分野における論文および統計学論文の引用状況について

張 菱軒[1]・潘 建興[2]・中野 純司[3,4]

（受付 2019 年 6 月 21 日；改訂 12 月 22 日；採択 2020 年 1 月 6 日）

要　　旨

ビッグデータ分析や機械学習などの出現により，統計学は近年大きな注目を集めている．そして学術論文においては，これまでもデータを正しく分析し新しい知見を裏付けるために統計学が広く使われてきた．ただ現代社会においては非常に多くの学術分野があり，それらの間の競争はますます激しくなっている．そのような競争の中で統計学が生き残るためには，他の学術分野論文への統計学論文の影響を客観的に測定することによって統計学の重要性を示すことが重要である．本研究では，各学術分野内の論文引用状況とそこでの統計学論文の引用状況を分析する．そのために学術論文データベース Web of Science を利用して学術分野を定義し，分析に必要となる引用数を集計した．

キーワード：論文引用解析，学術分野，Web of Science.

1. はじめに

統計学は，データの分析，解釈，生成過程のモデリングおよび推論を行う分野である．そのため，統計学はデータを用いるほとんどの分野で利用される．学術論文においては実験データの客観的な処理のために利用される．また，実用的にも広く利用されている．例えば，工学分野では，実験計画法や品質管理の統計手法を使って，生産物の品質の維持・改善が行われる．薬学では新薬の効果を測るために，臨床試験の結果を統計解析してその効果を分析する．マーケティングでは，A/B テストの結果が統計解析され，広告などの改良に利用される．金融分野や電子商取引の購入動向の分析のためには時系列解析がよく用いられ，予測や行動決定に利用される．

近年，研究分野間の競争は激しくなっており，そのため Institutional Research と呼ばれる分野で学問業績の客観的評価が重要になっている．したがって，統計学がどのように利用されているかを定量的に分析することは統計学にとって重要である．ただ，これまで他学術分野において統計学がどれくらい利用されているかを定量的に調べた研究は少ない．学術論文の影響を調べる一つの方法は，引用論文の状況を調べることである．当然のことながら，論文は自分と同じ分野の論文を多く引用する．論文の目的は新しい知見を発表することであり，そのために

[1] 総合研究大学院大学 複合科学研究科統計科学専攻：〒 190–8562 東京都立川市緑町 10–3
[2] 中央研究院 統計科学研究所：〒 11529 台北市南港區研究院路二段 128 號
[3] 中央大学 国際経営学部：〒 192–0393 東京都八王子市東中野 742–1
[4] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3

はこれまでその分野で知られていることを共有することが必要で，それには同分野の論文を引用しなければならないからである．しかしながら，異なる分野の論文が引用論文に含まれることも多い．その理由の一つは，同じようなテーマが複数の分野にわたって研究されていることがあるからである．特定のテーマの論文を検索すると，複数の分野の論文にたどり着くことがある．特に新しい分野はその傾向が強く，例えば機械学習は，コンピュータサイエンスと統計学の両方の知識が必要な手法であり，両分野の論文が機械学習の論文に多数引用される．このように，学術分野間には複雑な関係があり，その境界は曖昧であることも多い．

学術分野の引用情報の解析はこれまでもいくつか行われている．特に統計学分野の学術誌における引用情報の解析は Varin et al. (2016) で行われている．多くの学術分野に対しては，Leydesdorff (2004)，Zhang et al. (2010) のような解析が行われている．

本論文では，学術分野の論文が統計学論文を引用する頻度に着目する．それはその分野でどのくらい統計が必要とされるかを示すことになるからである．また，その値を指標として基準化するために分野内における引用状況も調べる．分野の分類と引用-被引用関係の集計を行うために，学術論文データベース Web of Science (WoS) を利用した．また解析のために，統計解析システム R (R Core Team, 2019) を使用した．

本論文の構成は以下の通りである．まず，2 節で WoS とそこでの分野分類について説明する．3 節で各分野における論文引用状況を調べる．次に 4 節で，各分野における統計学論文引用状況を調べる．最後に 5 節でまとめと注意を述べる．

2. Web of Science と学術分野

Web of Science (WoS, 2018) は，論文情報，著者情報，および学術誌情報などの学術情報に関する商用データベースであり，Clarivate Analytics 社により開発・維持されている．われわれは 1981 年から 2016 年までの WoS データを用いる．

WoS では各学術誌に 1 つまたは複数の分野が割り当てられている．ただし，140 の学術誌にはどの分野も割り当てられていない (例えば，International Review of Connective Tissue Research)．これらの分野のうち明らかな重複 (例えば Legal Medicine と Medicine, Legal) と学術誌名の明らかな重複 (例えば 2D Materials と 2D MATERIALS) を除く処理などを行うと，分野の数は 266 であり，全体で 19138 の学術誌と 45769924 の論文が含まれる．

割り当てられている分野数とそれに対する学術誌数および論文数は表 1 で示される．われわれが興味のある統計学は Statistics & Probability として分類されているが，これが単独で割り当てられている学術誌はなく，159 の学術誌がこれを割り当てとして含む．それに対して学術誌数と論文数を見ると，表 2 のようになる．Statistics & Probability を含む 2 つの分野が割り

表 1. 複数分野に割り当てられる学術誌数と論文数.

割り当てられる分野数	1	2	3	4	5	6	7	8	9	10
学術誌数	6792	6794	3062	1474	600	188	55	23	7	4
論文数	16270044	16674268	6618622	3818521	1540125	492942	183419	48700	25634	38414

表 2. 統計学に割り当てられる学術誌数と論文数.

割り当てられる分野数	2	3	4	5	6	7	8	9
学術誌数	74	20	17	14	13	1	3	2
論文数	104229	14890	26234	17769	20204	1158	13817	4997

当てられている 74 の学術誌はすべて Statistics & Probability と Mathematics に割り当てられている．なお，Mathematics だけに割り当てられている学術誌数は 220 である．

われわれは統計学分野に属する論文としては Statistics & Probability が割り当てられている学術誌に掲載された論文をすべて含めることにする．同様に他分野の論文として，その分野が含まれた学術誌に掲載された論文をすべて考える．そのため，例えば学術誌が 5 個の分野に割り当てられていれば，それに掲載された 1 つの論文が 5 個の分野で論文として集計される．

集計のために，われわれは 1981 年から 2016 年までの WoS データを用いて統計数理研究所で構築されたネットワークデータベースを利用する．このネットワークデータベースには 1981 年より前の論文は含まれていないので，例えば 1981 年の論文から引用される論文数はほとんど 0 と集計されることに注意する．

3. 各分野における論文引用特性

本節では各分野ごとの論文引用の状況をいくつかの指標を用いて分析する．最初に，各分野における 1 論文あたりの被引用数を見てみる．

$$\text{平均被引用数} = \frac{\text{被引用論文数}}{\text{総論文数}}$$

図 1 は分野ごとの平均被引用数を多い順に示したものである．図 1 は細部が見にくいため，図 2 では特に興味のあるランキングの上位 20 分野と下位 20 分野を拡大して示した．ひとつの論文が際立って多く引用されている分野は Astronomy & Astrophysics で平均 19.08 回引用されている．そのほかの分野では多くても 13.16 回である．Statistics & Probability の平均被引用数は 5.42 回である．この値はかなり少ないが，その理由は利用できるデータが 1981 年から 2016 年までのものなので，1980 年以前の論文を引用しても，それは引用回数としては数えられないからである．従って平均被引用数が上位の分野は同分野の論文を多く引用し，かつ新しい論文をよく引用し，古い論文をあまり引用しない分野と考えられる．下位 20 分野はほぼ文学・芸術関係の分野であり，平均被引用回数は 1 以下である．これらの分野では過去の論文を多く引用するか，論文をあまり引用しないと考えられる．

次に 1 回以上引用されている論文の割合を見る．

$$\text{1 回以上引用されている論文の割合} = \frac{\text{1 回以上引用されている論文数}}{\text{総論文数}}$$

図 3 は，分野における 1 回以上引用されている論文の割合を示し，図 4 はランキングの上位 20 分野，下位 20 分野を拡大して示したものである．この値は孤立していない論文の割合を示している．半分以上の分野で半分以上の論文が少なくとも 1 回引用されている．ここでも Astronomy & Astrophysics の値が最も高く，84% の論文が少なくとも一回引用されている．やはり文学・芸術関係の分野は孤立した論文が多いようである．

また，図 2 と図 4 の上位 20 分野を見ると，順位は少し異なっている．図 2 では，Neurosciences, Neurosciences & Neurology, Biochemistry & Molecular Biology, Virology, Geochemistry & Geophysics, Management と Psychology が上位 20 分野に入っているが，図 4 には入っていない．これは論文の引用頻度は多いが，それに較べて孤立している論文が多いことを示している．逆に図 4 では Chemistry, Inorganic & Nuclear, Parasitology, Polymer Science, Chemistry, Analytical, Fisheries, Oceanography と Materials Science, Biomaterials は上位 20 分野に入ったものの，図 2 には入っていない．これは論文の引用数の割には，孤立した論文が少ないと考えられる．

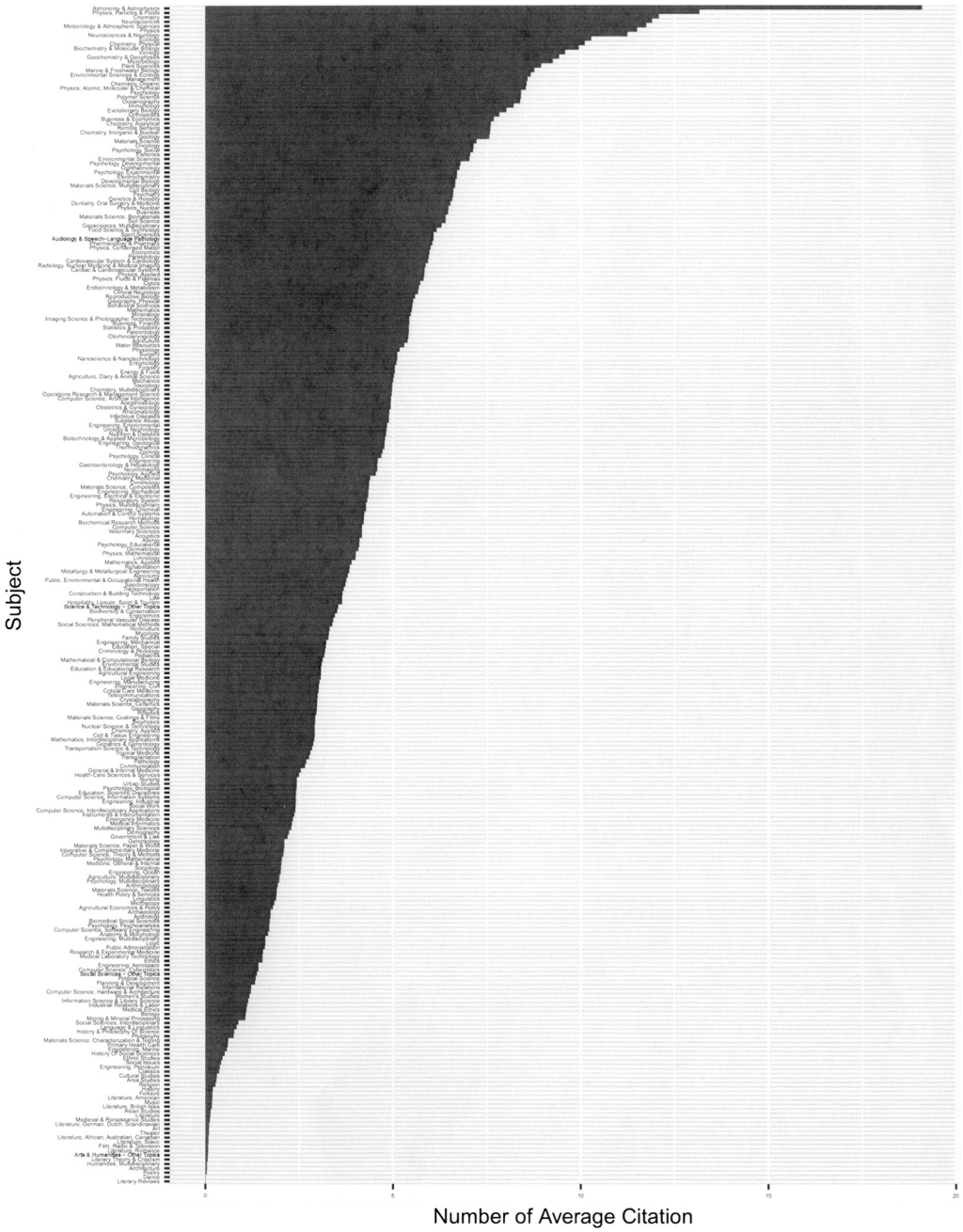


図 1. すべての分野における平均被引用数.

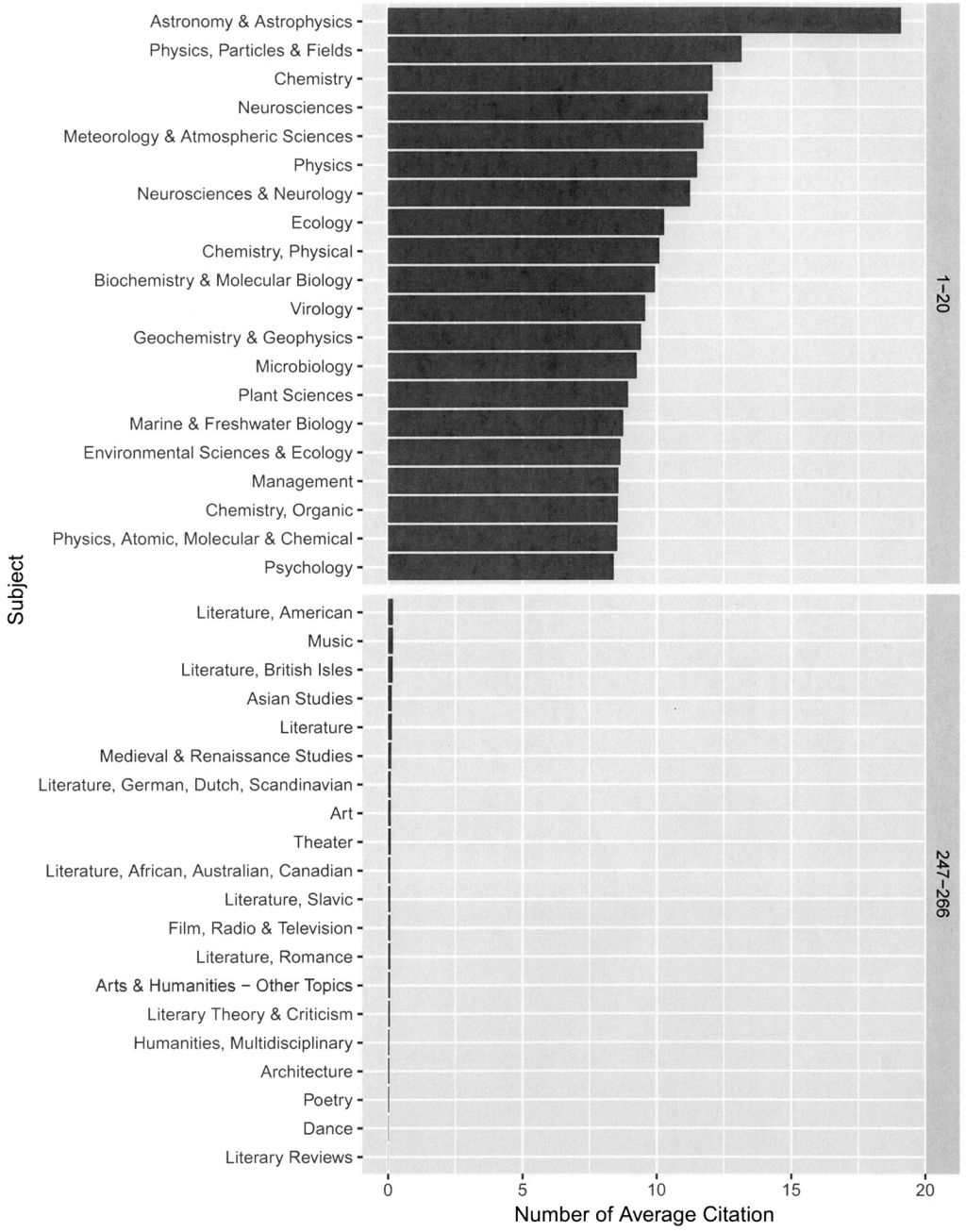


図 2. 平均被引用数の上位 20 分野と下位 20 分野.

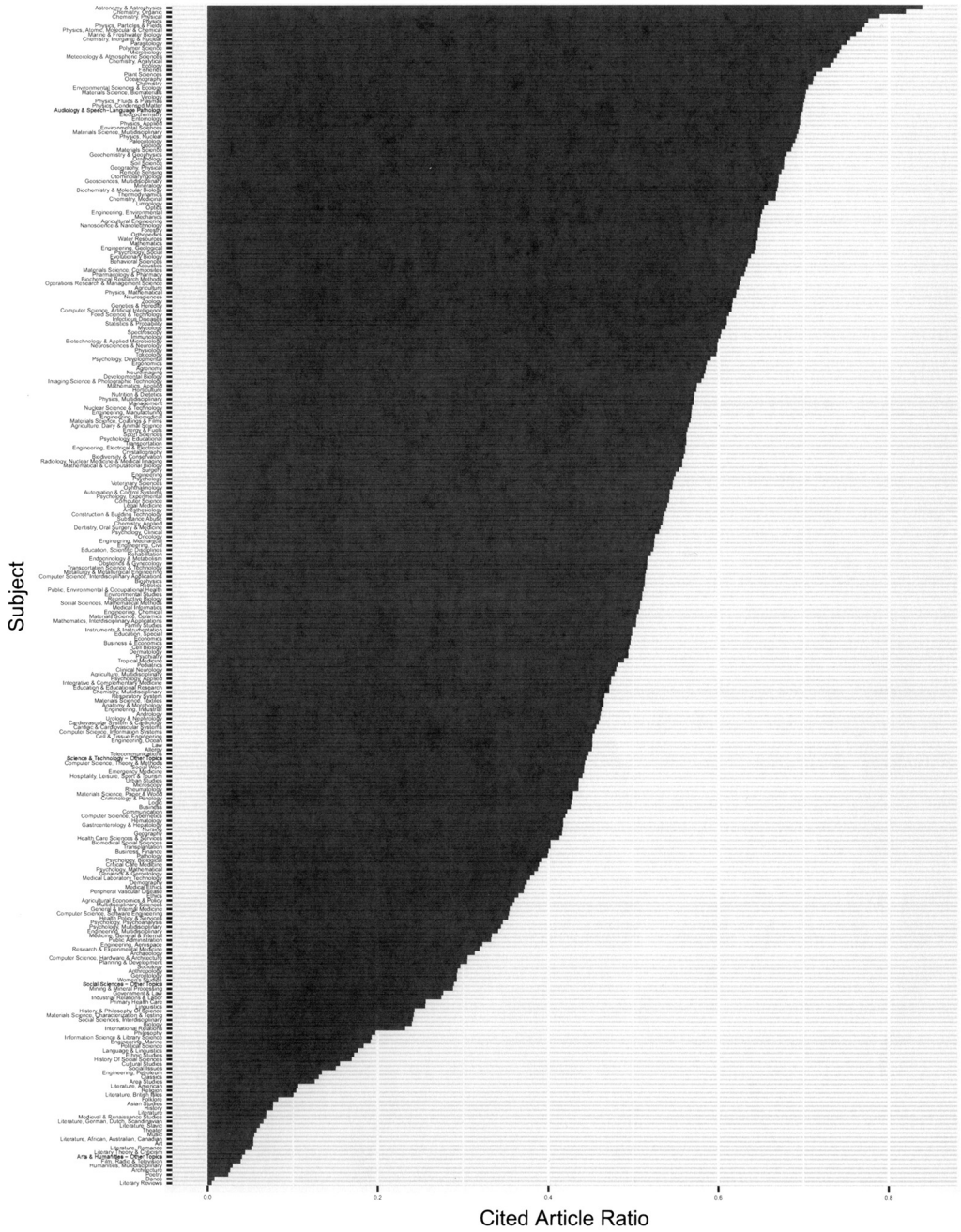


図 3. すべての分野における 1 回以上引用されている論文の割合.

図 4. 1 回以上引用されている論文の割合が上位 20 分野と下位 20 分野.

4. 各分野における統計学論文引用特性

最初に，各分野で 1 論文当たり何件の統計学論文が引用されているかを調べる．

$$\text{統計学論文平均被引用数} = \frac{\text{被引用統計学論文数}}{\text{総論文数}}$$

これはその分野で統計がどれくらい使われているかの一つのわかりやすい指標である．

図 5 は，すべての分野における 1 論文当たりの統計学論文被引用数である．図 6 はそのランキングの上位 20 分野，下位 20 分野を示した．Social Sciences, Mathematical Methods が最も多く統計学論文を引用している分野であり，一つの論文は平均的に 2 本以上の統計学論文を引用している．上位 20 分野の中では 6 分野が数学関連の分野である．下位 20 分野は文学・芸術関係の分野である．なお，近年注目されている Computer Science, Artificial Intelligence は 9 位であり，この分野で統計的手法がよく利用されていることがわかる．

しかしながら明らかに，この指標は各分野の平均論文被引用数に影響をうける．すなわち，もともと引用論文数が少ない場合は引用される統計学論文数も少ない．それで各分野の論文被引用数で基準化する．

$$\text{引用論文の中での統計学論文の割合} = \frac{\text{統計学論文被引用数}}{\text{論文被引用数}}$$

図 7 は，すべての分野における統計学論文引用率である．図 8 はランキングの上位 20 分野，下位 20 分野を示した．この場合，Mathematical & Computational Biology が 1 位となり，引用論文のうち約 80% が統計学論文である．また，図 6 の上位 20 位に入っていない Biology, Computer Science, Cybernetics, Research & Experimental Medicine, Industrial Relations & Labor が図 8 で上位 20 分野に入った．図 6 では上位 20 分野に入った Business &Economics, Evolutionary Biology, Automation & Control Systems と Mathematics, Applied は図 8 では入っていない．これらの分野では統計学論文を引用してはいるが，それは自分の分野の論文と比べるとそれほど多くないということを示す．

次に，どれくらいの異なる統計学論文が引用されているかを考える．すなわち分野ごとに，1 回以上引用された統計学論文数を総論文数で割って基準化した値を考え，図 9，図 10 に示す．

図 10 において，Computer Science, Cybernetics, Computer Science, Information Systems, Ergonomics と Management は上位 20 分野に入っているが，図 6 の上位 20 分野には含まれない．それはこの二つの分野では多くの異なる統計学論文を引用しているが，引用数自体はそれほど多くないことを示す．逆に，Economics, Business & Economics, Evolutionary Biology と Biochemical Research Methods は図 6 の上位 20 分野に入ったものの，図 10 では圏外になる．それは引用数は多いが，引用している論文の種類は多くないことを示しており，引用されている統計学論文が特定の論文に集中する傾向があることを示している．

次に基準化を行うときに，総論文数ではなくその分野で 1 回以上引用されている(孤立していない)論文数を利用する．この指標を考える理由は，孤立した論文はその分野における傍流あるいは例外と考えられるので，それらと統計学論文の関係を考慮する必要はないと思われるからである．

図 11，図 12 はその値を示す．ほとんどの分野では，同分野内の論文を引用することが多いのでこの指標の値は 1 より小さい．しかし，Social Sciences, Mathematical Methods だけは 1 より大きく，統計学論文の比重が大きいことがわかる．また，上位 20 分野において，やはり 6 分野が数学関連分野である．図 8 では上位に入っていない Biology, Demography と Computer Science, Software Engineering がここでは上位 20 分野に入っている．この分野では引用論文の

数が少ないが，その割には引用されている統計学論文の種類が多いことがわかる．図 10 では上位に入った Computer Science, Ergonomics と Management はここに入っていないので自分の分野を論文引用数と比べて，統計学論文の引用種類は多くないことを示す．また図 8 では上位に入っていない Automation & Control Systems, Demography, Computer Science, Information Systems, Computer Science, Software Engineering と Mathematics, Applied がここでは上位 20 分野に入っている．この分野では統計学論文の引用数は多くないが，引用する統計学論文の種類は多い事がわかる．逆に，図 8 では上位に入った Economics, Computer Science, Research & Experiment Medicine, Industrial Relations & Labor と Biochemical Research Methods は図 12 では入っていない．それは引用される統計学論文が集中する傾向があることを示す．

5. Conclusion

本論文では，最初に分野における引用状況を調べた．さらに特定の分野(Statistics & Probability)がその他の分野に引用される状況を調べた．その結果，統計学論文が他分野に与える状況が数値的に示された．

ここでは，統計学分野と他分野の関係を調べたが，同様の分析は他のすべての分野に関しても行える．特に数学，物理学のような基礎学術分野(統計学もその一つである)を分析することは興味深い．そして，このような分析は各分野が他の分野とどのように相互作用しているかを見ることにもなる．ここで計算したような指標は将来の共同研究の可能性や潜在的な異分野融合の指標として使用することもできる．明らかに，高い相互引用は分野間の強い関係を示すからである．また，論文の異分野融合の評価の指標として使う事も可能であろう．

謝　辞

本論文で使用したデータは Clarivate Analytics から提供されたものである．また，統計数理研究所の URA 室の本多啓介博士と濵田ひろか氏はデータを neo4j データベースに変換してくれた．それを使うことで本論文の分析が可能となった．栗木哲教授にはこの研究を遂行するためのよい環境を整えていただくとともに，研究上の助言もいただいた．査読者の方からは有益なコメントを頂いた．非常に感謝している．

参 考 文 献

Leydesdorff, L. (2004). Clusters and maps of science journals based on bi-connected graphs in journal citation reports, *Journal of Documentation*, **60**(4), 371–427, DOI: 10.1108/00220410410548144.

R Core Team (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria, https://www.R-project.org/.

Varin, C., Cattelan, M. and Firth, D. (2016). Statistical modelling of citation exchange between statistics journals, *Journal of the Royal Statistical Society*, *Series A*, **179**(1), 1–63.

Web of Science (2018). Clarivate Analytics, http://www.webofknowledge.com/.

Zhang, L., Glänzel, W. and Janssens, F. (2010). Journal cross-citation analysis for validation and improvement of journal-based subject classification in bibliometric research, *Scientometrics*, **82**(3), 687–706, DOI: 10.1007/s11192-010-0180-1.

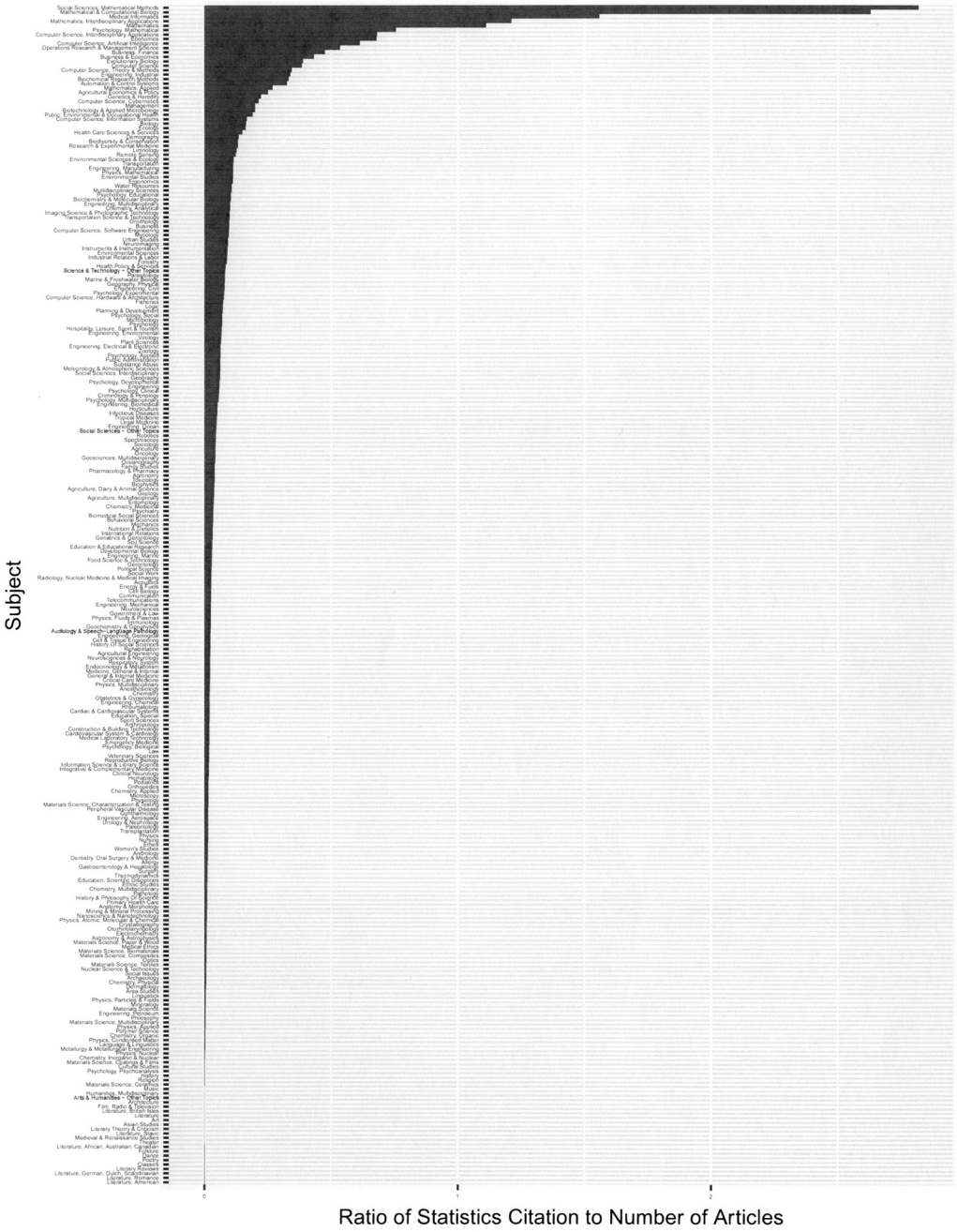


図 5. すべての分野における統計学論文平均被引用数.

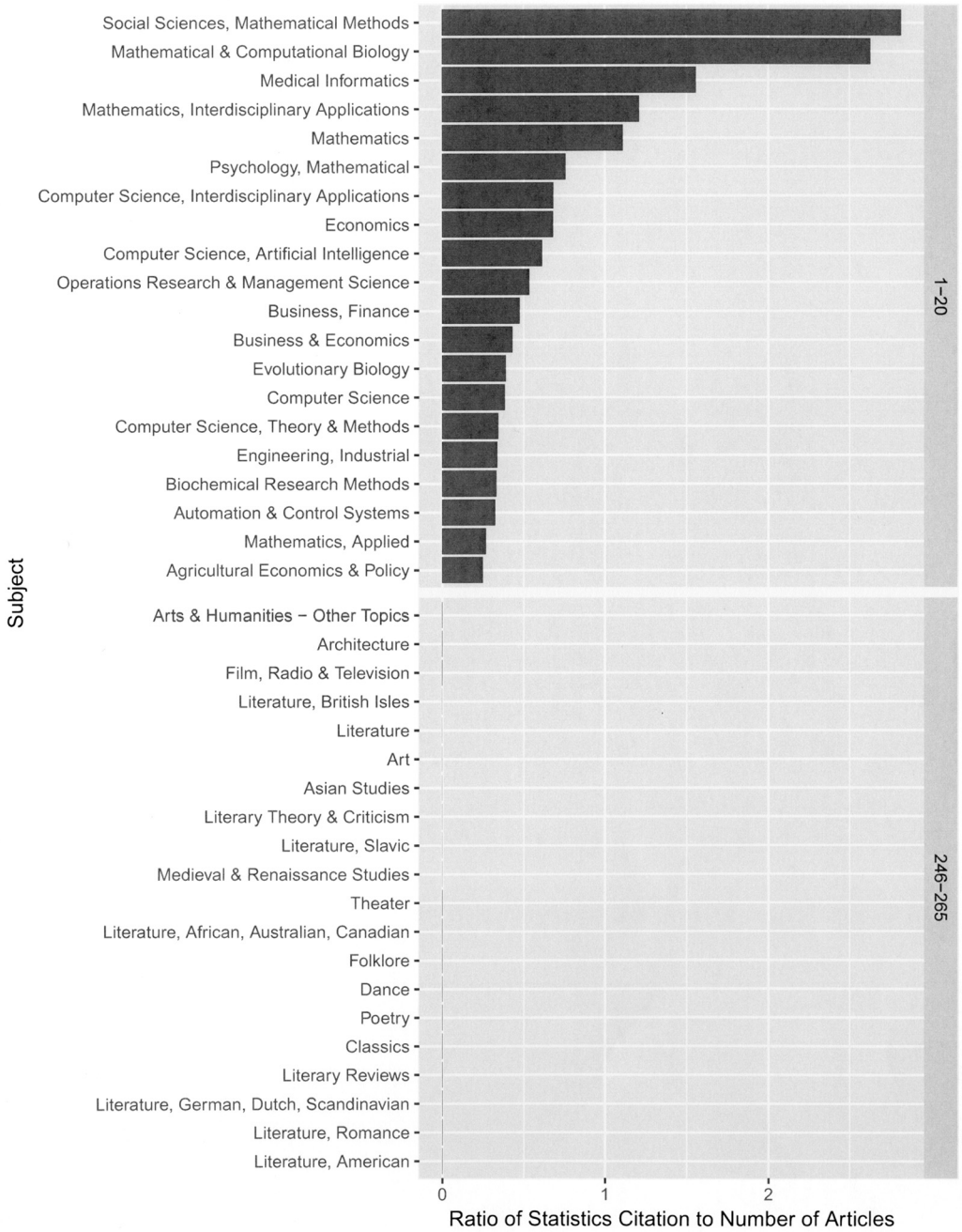


図 6. 統計学論文平均被引用数が上位 20 分野と下位 20 分野.

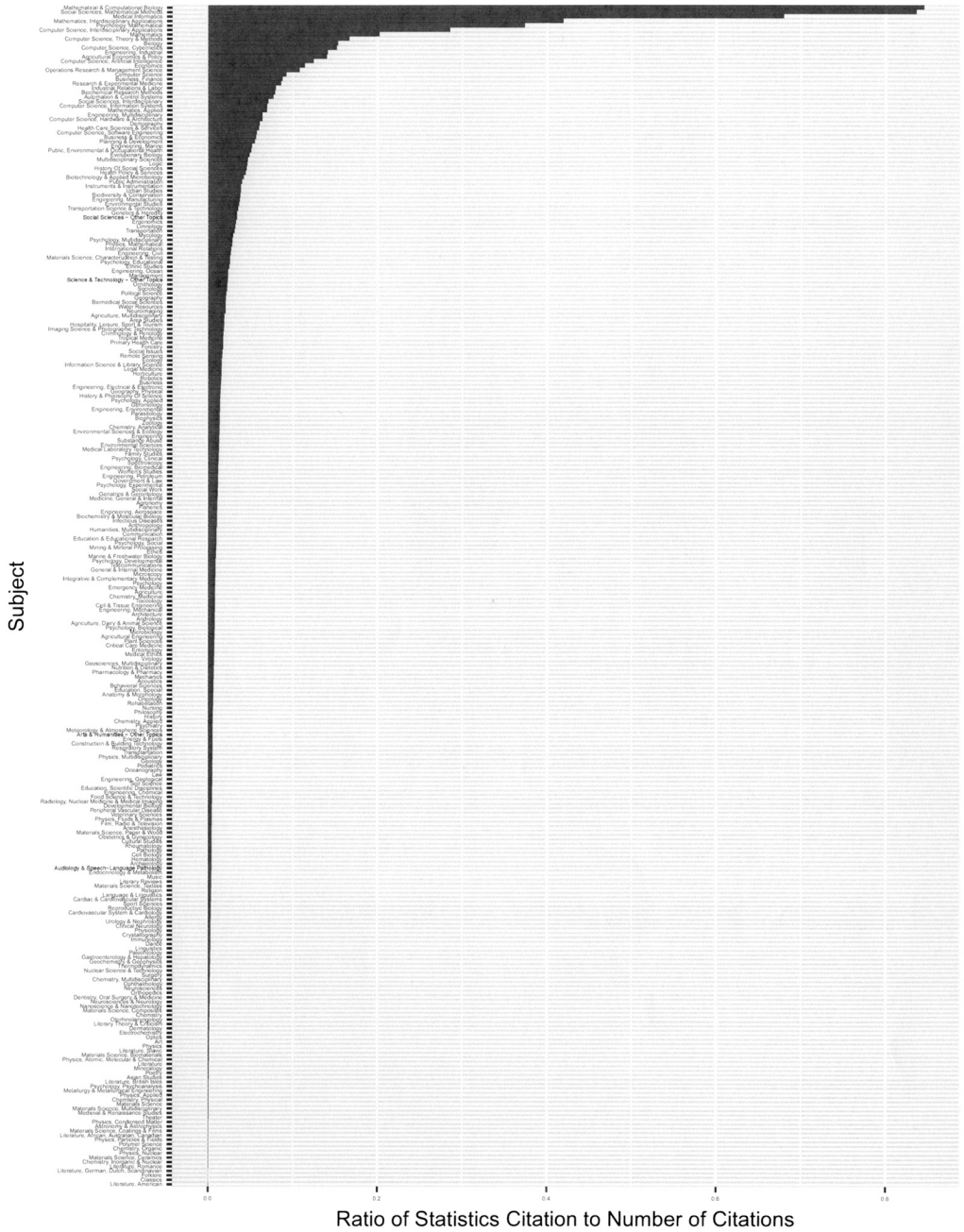


図 7. すべての分野における引用論文の中での統計学論文の割合.

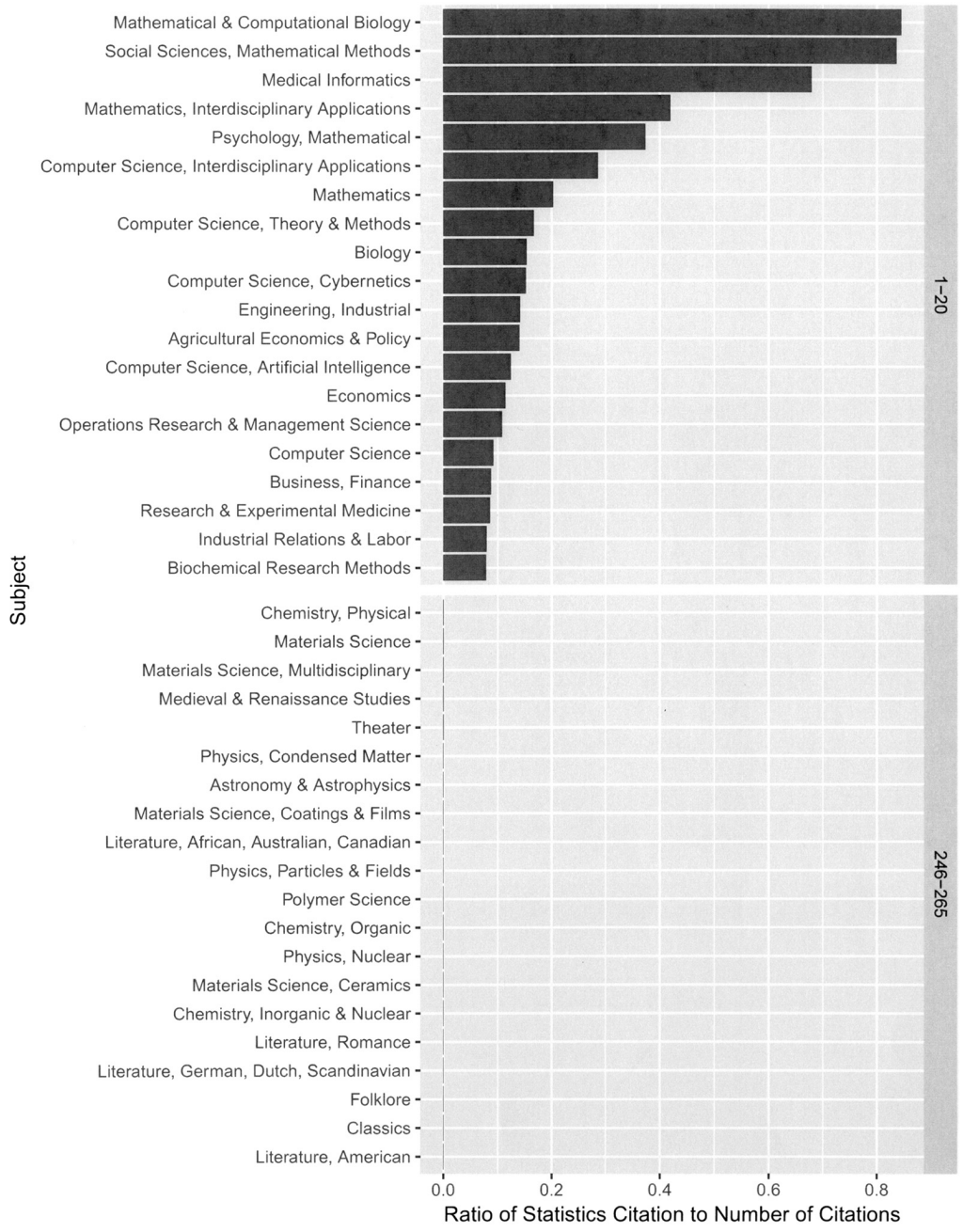


図 8. 引用論文の中での統計学論文の割合が上位 20 分野と下位 20 分野.

図 9. すべての分野における，1 回以上引用された統計学論文数(各分野の総論文数で基準化)．

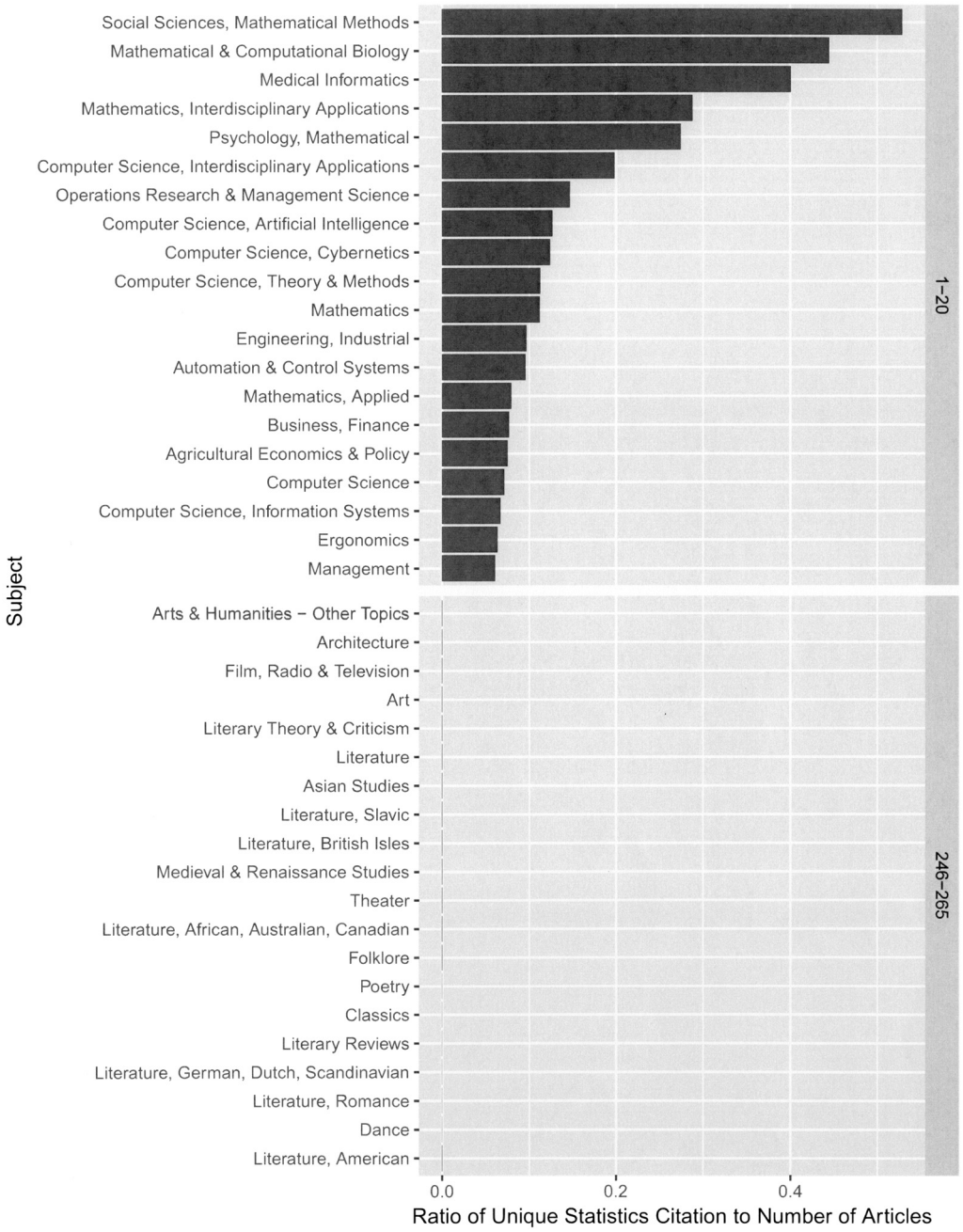


図 10. 1 回以上引用された統計学論文数(総論文数で基準化)が上位 20 分野と下位 20 分野.

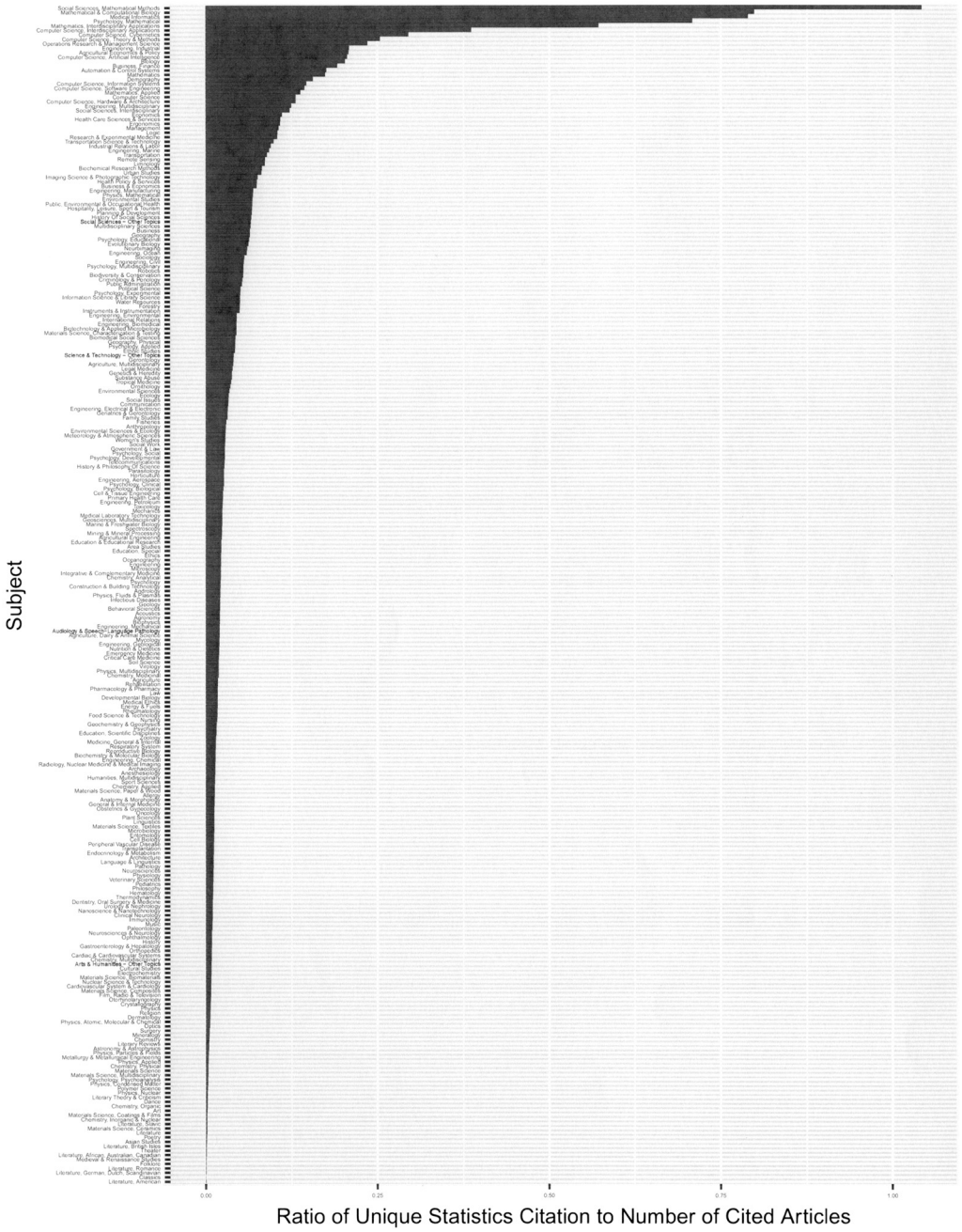


図 11. すべての分野における，1 回以上引用された統計学論文数(1 回以上引用されている論文数で基準化)．

図 12. 1 回以上引用された統計学論文数(1 回以上引用されている論文数で基準化)が上位 20 分野と下位 20 分野.

Citations of Academic Articles and Statistical Articles in Fields of Sciences

Livia Lin-Hsuan Chang[1], Frederick Kin Hing Phoa[2] and Junji Nakano[3,4]

[1]Department of Statistical Science, School of Multidisciplinary Sciences,
Graduate University for Advanced Studies
[2]Institute of Statistical Science, Academia Sinica
[3]Department of Global Management, Chuo University
[4]The Institute of Statistical Mathematics

Statistics has obtained more attention in recent years due to the rise of big data analysis and machine learning. Statistics are widely used in academic studies that require statistical analysis to objectively support their conclusions. In modern society, there exist many academic fields, and competition among them is severe. In order for statistics to survive such competitions, it is important for statisticians to measure the influence of articles in the field of statistics relative to those in other academic fields. In this work, we analyze citations within each academic field, focusing on citations of statistical articles. We used a database of academic articles from "Web of Science" to define academic fields and to count the required numbers of citations in the study.

Key words: Academic fields, citation analysis, Web of Science.

統計数理(2020)
第 68 巻 第 2 号 265–285

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

学術文献DBを用いた共著分析によるIoT研究における異分野融合の国際比較

水上 祐治[1]・中野 純司[2,3]

(受付 2019 年 11 月 5 日；改訂 2020 年 6 月 2 日；採択 6 月 3 日)

要　　旨

2011 年，ドイツ技術科学アカデミーとドイツ連邦教育科学省は，「あらゆる社会システムの効率化」「新産業の創出」「知的生産性の向上」を目指した技術的フレームワーク Industry 4.0 を発表した．Industry 4.0 は，サイバーフィジカルシステムというコンセプトのもと，IoT 技術，Big-Data 技術，人工知能技術を駆使して，現実世界(フィジカル空間)での現象を膨大な観測データとして蓄積，サイバー空間の強力な計算資源と結びつけて意思決定に活用するものである．Industry 4.0 発表後，それら要素技術の研究が世界的に活発化している．本稿は，フィジカル空間とサイバー空間の橋渡しをする IoT 技術の研究に着目，異分野融合による研究促進の視点から，各国の研究促進の戦略を分析して考察するものである．分析では，一連の研究で導出された分析手法と共に，主成分分析，階層型クラスター分析を行い，多角的に考察を展開した．分析の結果，IoT 研究の異分野融合では，「工学とコンピューター科学の連携比率の高さ」という技術優先的アプローチ，そして，「化学と臨床医学の連携比率の高さ」というアプリケーション的アプローチにより，IoT 論文上位 10 か国を 3 つのグループに分類することができた．

キーワード：IR，研究力評価，異分野融合，共著分析，イノベーション．

1. Industry 4.0 と主要技術

2011 年，ドイツ技術科学アカデミーとドイツ連邦教育科学省は，Industry 4.0 のフレームワークを発表し，2013 年に最終報告書をまとめた．Industry 4.0 は，「あらゆる社会システムの効率化」「新産業の創出」「知的生産性の向上」を目指した技術的フレームワークであるが，特に Cyber Physical System(CPS) の概念に基づいた先進的な工場(Smart Factory)の普及を促進することで，工場の生産活動の効率を改善することを目指している (Putnik et al., 2019)．なお，工場の生産活動の効率を高める活動は以前から試みられていたが，Industry 4.0 の特徴は，事前に機器の故障や異常を予測して予防する「予測メンテナンス」にある．IoT(Internet of Things) 技術，Big-Data 技術，AI(Artificial Intelligence：人工知能)技術は，このような特徴を生み出す手法として注目されている．図 1 に，Industry 4.0 におけるこれら技術の位置づけを示す．

IoT 技術は，センサーを機器に組み込み，インターネット経由で機器情報を送信する技術の総称である．Big-Data 技術は，IoT 技術等によって収集された大量のデータを整理および保存

[1] 日本大学 生産工学部：〒 275–8575 千葉県習志野市泉町 1–2–1
[2] 中央大学 国際経営学部：〒 192–0393 東京都八王子市東中野 742–1
[3] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3

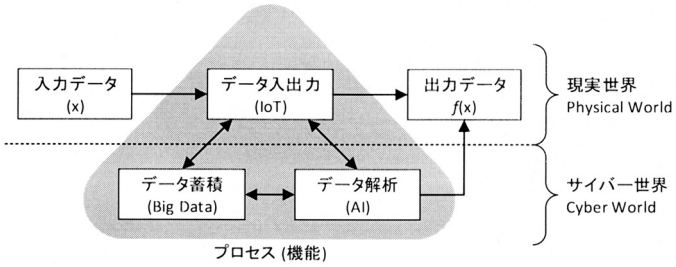


図 1. Industry 4.0 における IoT，AI，Big-Data の位置づけ．

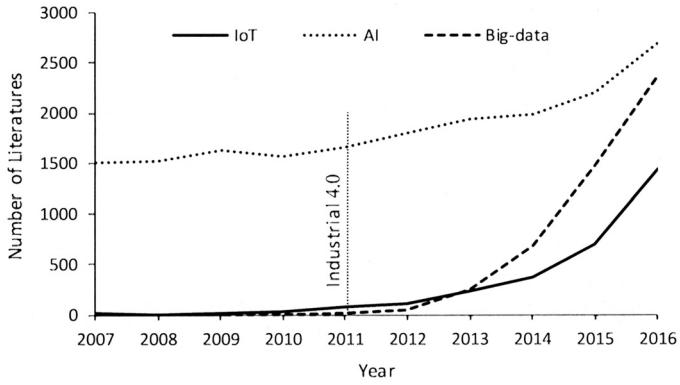


図 2. IoT，AI，Big-Data の論文数の推移．（この調査には，Web of Science Core Collection を用いた．WoS のサーチワードは，IoT 技術では，<Search word [IoT]: (TS=(IoT) OR TS="Internet of Things") AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article)>である．AI 技術は，<Search word [AI]: (TS=(AI) OR TS="Artificial Intelligence") AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article)>である．最後に，Big-Data 技術は，<Search word [Big-Data]: (TS="Big Data") AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article)>である．）

するための技術の総称である．AI 技術は，Big-Data 技術によって蓄積された大量の情報を分析するための技術である．

1.1 IoT 技術，Big-Data 技術，AI 技術の論文数の推移と国別比較

Industry 4.0 の普及に伴い，IoT 技術，Big-Data 技術，AI 技術に関する研究が活発に行われている．図 2 に，これら技術に関する論文数の年次変化を示す．

IoT，Big-Data，AI の論文数は共に継続的に増加しており，2016 年には，IoT が 1439 編，Big-Data が 2369 編，AI が 2704 編に達している．AI の論文数は，2007 年の時点で年間 1500 編であり，これら技術の中では際立って多い．AI は，現在，3 次ブームにあるといわれており，1980 年代から論文が蓄積され，2007 年以降も継続的に増加している．次に，IoT と Big-Data の論文数であるが，2011 年頃に Industry 4.0 が発表される前後で，その傾向が変化している．まず，2011 年以前では，IoT，Big-Data 共に論文数は毎年数十件にとどまっていた．しかし，2011 年頃に Industry 4.0 が発表されてからは，IoT，Big-Data 共に論文数が増加し，特に 2013

表 1. IoT/AI/Big-Data の論文数上位 10 か国（2016）.

順位	国・地域		
	IoT	AI	Big-Data
1	中国	アメリカ	アメリカ
2	アメリカ	中国	中国
3	韓国	イギリス	イギリス
4	イギリス	インド	オーストラリア
5	イタリア	ドイツ	ドイツ
6	スペイン	イラン	カナダ
7	日本	ブラジル	韓国
8	インド	スペイン	イタリア
9	ドイツ	フランス	スペイン
10	台湾	カナダ	日本
		日本(13)	

年以降は増加率が急激に高まっている.

表 1 に，IoT，Big-Data，AI 論文の国・地域別での順位を示す．中国，アメリカ，英国においてこれらの分野の論文が多く生み出されていることが示されている．一部の分野に特化している国・地域があり，IoT では韓国が 3 位，AI ではインドが 4 位，Big-Data では，オーストラリアが 4 位に入っている．一方，日本の順位は，IoT，Big-Data，AI の各分野でそれぞれ 7 位，13 位，10 位であった．なお，イギリスは，イングランド，スコットランド，北アイルランドを含む地域とする.

2. 本稿の目的

本稿は，フィジカル空間とサイバー空間の橋渡しをする IoT 技術の研究に着目，異分野融合によるイノベーション誘発と研究促進の視点から，各国の研究促進の戦略を分析して考察することを目的としている．イノベーションの概念は，いくつかの側面があり，現状では，その定義は定まっていない．しかし，一側面として，イノベーションは，異なる分野の融合(異分野融合)がもたらす新たな価値と定義することができる（Mizukami et al., 2016）．学術研究の分野における異分野融合の類型として，「異なる組織間の共同研究」，「異なる研究分野間の共同研究」，「産学官連携による共同研究」などが挙げられる．本稿では，IoT 技術の研究における「異なる研究分野間の共同研究」に着目して，イノベーション戦略の側面から各国の研究促進の戦略を分析して考察するものである.

なお，現状の把握という目的では，直近の年のデータを分析するのが望ましいが，書誌データベースの特性上，直近数年は，データが更新されていないなど，ノイズが多い場合があるため，データが安定していると考えられる 2016 年を対象年とした.

3. 関連分野のレビュー

3.1 研究評価の分類

書誌データを用いた研究評価の方法には，文献の発表数を対象に分析する論文数調査の系統と，論文への引用度数や論文間の引用による結合関係を対象に分析する引用統計・分析の 2 系統がある（根岸・山崎, 2001; Wagner et al., 2011）．研究評価の分類を表 2 に示す．前者の「論文数調査」の目的は，論文の生産性指数，研究活動の規模指数を求めることであり，主に単純集計の手法が用いられる．論文数調査では，分野，年，国・地域，および，所属，さらに，他の経済統計指標との相関分析が実施されている(根岸・山崎, 2001)．たとえば，Vergidis et al. (2005)の微生物学における論文生産性の国際比較がある．Vergidis et al. (2005)は，1995〜2003

表 2. 研究評価の分類.

系統	目的	分析手法	
論文数調査	論文の生産性指数 研究活動の規模指数	単純集計	
引用統計・分析	論文の消費指数 研究活動の品質指数	引用分析 共著分析 (謝辞分析)	(共語分析) (共分類分析)

年において，微生物学における論文生産性は，西ヨーロッパが最も高い成長率を示して，続いて，北米，アジア，中南米，および，東ヨーロッパの順であるとしている．

一方，後者の「引用統計・分析」の目的は，論文の消費指数，研究活動の品質指数を求めることであり，引用分析，共著分析等の手法が用いられる（根岸・山崎, 2001; 藤垣 他, 2004）．また，少数ではあるが，謝辞分析，共語分析，共分類分析等の手法が用いられる（藤垣 他, 2004）．なお，謝辞分析は，論文中に記されている謝辞が分析対象，そして，共語分析は，論文中に記されている複数の語の間の関係が分析対象，最後に共分類分析は，分野分類の共出現現象が分析対象である（藤垣 他, 2004）．また，その他の指標として多様性分析がある．多様性分析は，Rafols and Meyer（2010），Stirling（2007），Porter and Rafols（2009）によって提案されている．Rafols and Meyer（2010）は，事前定義されたカテゴリを用いて，計量書誌データの多様性を記述する目的にて多様性指標を提示している．

本稿では，「引用統計・分析」の共著分析を用いる．共著分析は，個人の業績審査に適応可能であり，例えば，協力関係にある機関の研究員との共同研究がどの程度行われているかを把握することが可能である（藤垣 他, 2004）．

3.2 異分野融合と研究力のイノベーション

学術研究の分野における異分野融合の類型として，「異なる組織間の共同研究」，「異なる研究分野間の共同研究」，「産学官連携による共同研究」などが挙げられる．

「異なる組織間の共同研究」に関する関連研究として，Mizukami et al.（2016）が挙げられる．この研究では，組織の研究力を高めるためにはイノベーションを起こすことが有効であるとの前提のもと，組織内外の共同研究が重要であるとして，論文共著情報をもとにしたそれら協力関係の測定手法を提案した．当該手法では，ネットワーク理論の媒介中心性指標の概念を組織論に適用できるように拡張して，組織内，組織外，組織内外のつながりを個別に集計することが可能となり，組織内外の情報の流れを管理して，イノベーションの起こりやすい組織を目指すことが可能となった．

また，「異なる組織間の共同研究」に関する国際比較として，Mizukami et al.（2017）が挙げられる．この研究では，これら媒介タイプの比率を個人別に算出して国別で集計，その分布をローレンツカーブとジニ係数で表して，媒介タイプとイノベーションの普及との関連について考察を行った．さらなる「異なる研究分野間の共同研究」に関する関連研究として，Mizukami et al.（2018）が挙げられる．この研究では，研究者の専門分野を客観的に定義することを目指し，共著分析をもとにした専門分野の導出手法を提案した．本稿では，「異なる研究分野間の共同研究」に関して共同研究が活発に行われている分野を把握する方法を提案する．「産学官連携による共同研究」に関して，文部科学省は「研究力向上改革 2019」（2019）において，その重要性を示している．

4. 本稿の分析手法

4.1 研究者の専門分野の特定

従来，研究者の専門分野は，各個人の申請に基づくことがあり，主観的な定義であった．また，主観的な定義であるので，専門分野での実績が伴わないことがあり，研究者の専門分野の客観性が明確でない問題があった（Mizukami et al., 2017）．

この問題に対して，Mizukami et al.（2017）は，研究者の専門分野を客観的に定義することを目指し，共著情報をもとにした専門分野の導出手法を提案した．図 3 に研究者 A の専門分野とその応用分野の一例を示す．研究者 A の発表論文は，数学分野①②が 2 編，臨床医学分野③が 1 編，経済学 & ビジネス分野⑤が 1 編，総合分野①④が 1 編であった場合，研究者 A の専門分野は，数学分野であり，その集中度は，40.0% であるとするものである．集中度が高い場合，その研究者は，専門分野の研究に集中していると考えられる．一方，集中度が低い場合，その研究者は，専門分野の研究成果を他の分野に応用していると考えられる．

表 3 に本稿で用いた研究分野の分類を示す．なお，この分類は，Clarivate Analytics の Web of Science Core Collection に掲載の Essential Science Indicators Subject Areas（ESISA）を元にしている．

4.2 組織の研究力と異分野融合度の見える化

組織の研究力と異分野融合度を見える化する手法は，図 3 で示した各研究者の専門分野とその応用分野の情報を元に，組織に所属する研究者の情報を重ねて，組織の研究力と異分野融合

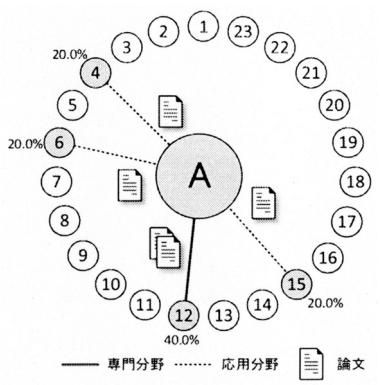


図 3. 研究者 A の専門分野とその応用分野の一例.

表 3. 本稿の研究分野の分類.

番号	研究分野	番号	研究分野	番号	研究分野
1	農学	9	地球科学	17	薬理学 & 毒物学
2	生物学 & 生化学	10	免疫学	18	物理学
3	化学	11	物質科学	19	植物 & 畜産学
4	臨床医学	12	数学	20	心理学/精神医学
5	コンピューター科学	13	微生物学	21	社会科学・一般
6	経済学 & ビジネス	14	分子生物学 & 遺伝学	22	宇宙科学
7	工学	15	総合	23	人文科学
8	環境/生態学	16	神経科学 & 行動		

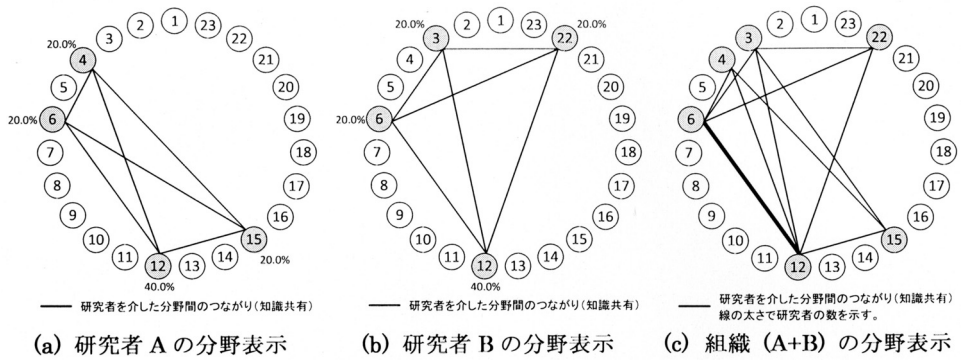


図 4. 研究者の分野表示と組織の分野表示の一例.

を示すものである．しかし，図 3 で示した各研究者の情報は，中央に示した各研究者を介さないと各研究分野のつながりを示すことができず，そのつながりが認識しづらい．そこで，当該手法では，まず，各研究者の情報を分野間の情報のみに組み替えた簡易版の研究者の分野表示手法を用いる．図 4(a) に簡易版の研究者 A の分野表示手法の一例を示す．例えば，臨床医学分野③と数学分野①②が研究者 A を介してつながり知識が共有されている．このように，図 4(a)では，研究者 A を介した各分野のつながりが明確に示されている．

次に組織に所属する研究者の情報を重ねて，組織の研究力と異分野融合を示す．図 4(c)は，図 4(a)の研究者 A と図 4(b)の研究者 B を重ねて，組織の研究力と異分野融合度を示したものである．図 4(c)のビジネス分野⑤と数学分野①②の間の太い線は，研究者 A と研究者 B の 2 名を介したつながりである．その他の分野間の細い線は，研究者 A または研究者 B のどちらか 1 名を介したつながりである．このように，図 4(a) では，図 4(a)の研究者 A と図 4(b)の研究者 B を介した各分野のつながりが明確に示されている．組織の研究力と異分野融合度を見える化する手法では，組織間のつながりにおいて，媒介する研究者の数が多い，または，比率が高い場合にそれら分野間の知識の共有が進むとの理解のもと接続線の幅を太く示している．

5. 分析

5.1 分析の手順

本稿で提案する分析，「組織の研究力と異分野融合の分析」，「異分野融合の類似性に着目した国別の階層型クラスター分析」，「異分野融合の分類の要因を定量的に把握するための主成分分析」の 3 段階で構成されている．まず，図 5 に組織の研究力と異分野融合の分析手順を示す．この分析では，分析対象の分野を確定することから始まる．

図 5 では分野 A とした(❶)．次に，分野 A の論文収集を行う(❷及び❷A)．続いて，分野 A の著者の専門分野を特定するために著者が関わる論文を全て収集する(❸)．ここでは，前工程(❷)で収集した論文群から，その全著者を抽出(❸A)して，その著者が当該年に執筆した全論文の収集を行った(❸B)．次に，著者の専門分野特定を行った(❹及び❹A)．この段階にて，各研究者が関わる専門分野がすべて明らかになる．最後に，組織の研究力の特定と異分野融合度の算出を行う(❺及び❺A)．

次の「異分野融合の類似性に着目した国別の階層型クラスター分析」では，253 種類の分野間つながりに対して，階層型クラスター分析を適用し，デンドログラムで可視化する．最後に「異分野融合の分類の要因を定量的に把握するための主成分分析」では，253 種類の分野間つな

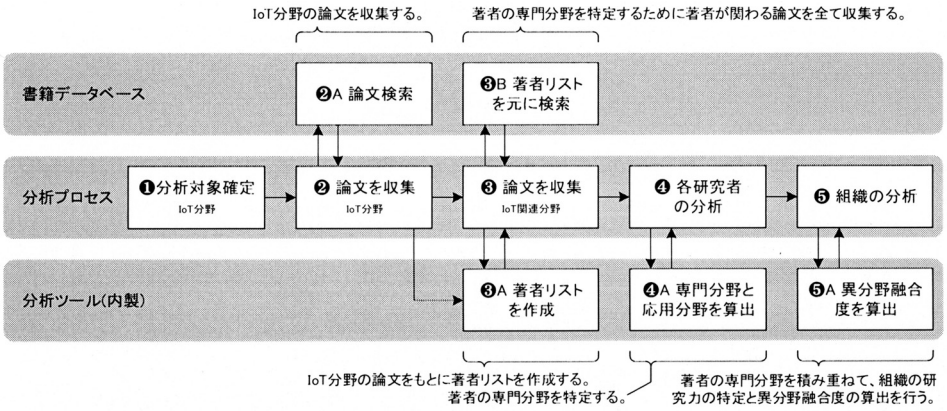


図 5. 組織の研究力と異分野融合の分析手順.

表 4. IoT 論文数上位 10 か国の地域別論文数 (2007–2016). #1：比率の平均は調和平均とする.

順位	国・地域	論文				
		数量	地域別比率 [%]			
			全地域	EU	NA&SA	Asia
1	中国	825	23.285			23.285
2	アメリカ	520	14.677		14.677	
3	韓国	324	9.145			9.145
4	スペイン	237	6.689	6.689		
5	イギリス	234	6.605	6.605		
6	イタリア	217	6.125	6.125		
7	台湾	140	3.951			3.951
8	ドイツ	134	3.782	3.782		
9	フランス	122	3.443	3.443		
10	日本	115	3.246			3.246
-	その他	675	19.052			
上位 10 か国の平均		354.300	5.554#1	4.907#1	14.677#1	5.606#1
-	合計	3543	100.000	26.644	14.677	39.627

がりに対して，主成分分析を適用し，分類要因の抽出および国別の特徴の定量化を行う．なお，データの収集元は Web of Science Core Collection，著者抽出には Microsoft Visual Basic for Applications(VBA)による内製の分析ツール，主成分分析とクラスタリング分析には統計解析ソフトウェア R を用いた.

5.2 収集データ

Industry 4.0 の要素技術である IoT，Big-Data，AI の論文数は共に継続的に増加しており，2016 年には，IoT が 1439 編，Big-Data が 2369 編，AI が 2704 編に達している．本稿では，フィジカル空間とサイバー空間の橋渡しをする IoT 技術に関して分析を行いその詳細を明らかにする．表 4 に 2007 年から 2016 年までの 10 年間における IoT 関連論文の国別比較を示す．なお，NA&SA は，北米と南米を含むものとする.

全体の論文数は 3543 編であった．国別では，論文数 1 位は，中国 825 編で全体の 23.285% であった．2 位以下は，アメリカ 520 編(14.677%)，韓国 324 編(9.145%)であった．20% を超えるのは中国のみであり，アメリカが 10% 以上で，上位 3 か国で全体の 47.107% を占めている．そして，スペイン，イギリス，イタリアが 6% 台，台湾，ドイツ，フランス，日本が 3% 台で

表 5. IoT 論文の研究分野（2007–2016）.

番号	研究分野	論文数	比率 [%]	番号	研究分野	論文数	比率 [%]
1	農学	6	0.169	14	分子生物学 & 遺伝学	11	0.310
2	生物学 & 生化学	15	0.423	15	総合	39	1.101
3	化学	230	6.492	16	神経科学 & 行動	9	0.254
4	臨床医学	34	0.960	17	薬理学 & 毒物学	0	0.000
5	コンピューター科学	1301	36.720	18	物理学	13	0.367
6	経済学 & ビジネス	61	1.722	19	植物 & 畜産学	4	0.113
7	工学	976	27.547	20	心理学/精神医学	6	0.169
8	環境/生態学	12	0.339	21	社会科学・一般	96	2.710
9	地球科学	21	0.593	22	宇宙科学	2	0.056
10	免疫学	1	0.028	23	人文科学	12	0.339
11	物質科学	21	0.593	-	指定なし	664	18.741
12	数学	9	0.254	-	合計（指定なしを除く）	2,879	81.259
13	微生物学	0	0.000	-	合計	3,543	100.000

あった．次に地域別では，上位 10 か国において，欧州連合が 5 か国で 26.644%，北米＆南米がアメリカのみで 14.677%，アジアが 4 か国で 39.627% であった．

表 5 に，論文の研究分野別の論文数と比率を示す．1 位は，コンピューター科学 1301 編で全体の 36.720% であった．以下，工学 976 編（27.547%），化学 230 編（6.492%）で，上位 3 分野で全体の 70.759% を占めている．なお，全体の論文数は 3543 編であったが，分野指定に欠損があるものがあり分野が確定できたものは 2879 編であった．

5.3 分析方法と分析結果

5.3.1 IoT 分野の論文を収集と著者の抽出

ステップ❶❷では，IoT 分野の論文を収集する．検索条件は，トピックが「IoT」または「Internet of Things」であり，ドキュメントタイプが「Article」または「Review」である 2016 年の英語ドキュメントである．分析対象は，一般的には最新のデータを分析することが望ましい．しかし，書誌データベースの特性により，直近数年間は，論文の更新が頻繁に行われるため，採録論文の網羅性が低い可能性がある．したがって，本稿では，直近で網羅性が高いと考えられる 2016 年を分析対象とした．ステップ❸では，IoT 分野の論文をもとに IoT 分野の著者の洗い出しを行う．抽出の結果，2016 年の IoT 分野では論文数が 1663 編であり，それら論文から，累積で 6028 名の著者が抽出された．表 6 に，2016 年の IoT 分野における論文数と著者数を示す．なお，NA&SA は，北米と南米を含むものとする．

国別では，論文数 1 位は，中国 344 編で 1278 名であり，著者数で全体の 21.201% であった．以下，アメリカ 246 編（919 名，15.246%），韓国 193 編（598 名，9.920%）であった．著者数比率で 20% を超えるのは中国のみであり，アメリカが 15% 以上で，上位 3 か国で全体の 46.367% を占めている．次に地域別では，上位 10 か国において，欧州連合が 4 か国で 26.775%，北米＆南米が 15.246%，アジアが 4 か国で 41.573% であった．その他の特徴として，著者数と論文数の比率では，欧州連合のイギリス，イタリア，スペイン，ドイツが平均以上の値となっている．

5.3.2 IoT 分野の著者および IoT 関連分野の論文の分布

ステップ❹では，著者の専門分野を特定する．ステップ❸で作成された著者リストに基づいて，著者の 2016 年に公開された全ての論文を収集する．表 7 に国別の著者の 2016 年に公開された全ての論文を示す．

2016 年の IoT に関する論文の著者 6028 名に対して 154353 件の論文が抽出された．研究分野別では，1 位は，3（化学）の 29194 編で 18.914% であった．以下，4（臨床医学）28605 編

表 6. IoT 論文数上位 10 か国の地域別論文数と著者数（2016）．#1：比率の平均は調和平均とする．

順位	国・地域	論文	著者/論文	著者				
		数量	数量	数量	地域別比率［%］			
					全地域	EU	NA&SA	Asia
1	中国	344	3.715	1278	21.201			21.201
2	アメリカ	246	3.736	919	15.246		15.246	
3	韓国	193	3.098	598	9.920			9.920
4	イギリス	110	4.500	495	8.212	8.212		
5	イタリア	108	4.157	449	7.449	7.449		
6	スペイン	84	4.702	395	6.553	6.553		
7	日本	77	3.247	250	4.147			4.147
8	インド	59	3.068	181	3.003			3.003
9	ドイツ	58	4.741	275	4.562	4.562		
10	台湾	55	3.618	199	3.301			3.301
-	その他	329	3.006	989	16.407			
上位 10 か国の平均		133.400	3.858	503.900	5.819#1	6.371#1	15.246#1	4.878#1
-	合計	1663	-	6028	100.000	26.775	15.246	41.573

表 7. IoT 関連論文における著者の全論文（2016）．

研究分野	合計	比率［%］	1 中国	2 アメリカ	3 韓国	4 イギリス	5 イタリア	6 スペイン	7 日本	8 インド	9 ドイツ	10 台湾
1	2842	1.841	1468	525	275	355	17	6	61	50	19	66
2	4999	3.239	2411	1040	482	683	43	19	109	31	49	132
3	29194	18.914	15098	5602	2447	3700	238	95	754	152	210	898
4	28605	18.532	13179	5888	3316	3793	283	121	771	145	275	834
5	7420	4.807	2882	1485	785	930	301	180	263	194	110	290
6	1059	0.686	432	203	91	166	41	7	31	26	37	25
7	23024	14.916	10825	4409	2194	2952	383	246	657	261	248	849
8	4542	2.943	2312	862	330	619	40	38	111	28	48	154
9	3806	2.466	1935	777	258	511	34	26	96	22	35	112
10	811	0.525	415	151	71	108	7	5	18	9	9	18
11	8594	5.568	4402	1649	750	1123	52	9	227	61	54	267
12	1432	0.928	718	267	91	207	26	15	26	13	17	52
13	1424	0.923	729	271	122	158	13	3	57	9	14	48
14	9434	6.112	4778	1863	844	1212	76	16	216	49	67	313
15	8815	5.711	4313	1788	842	1127	71	21	247	58	74	274
16	2480	1.607	1167	528	263	303	34	10	70	7	37	61
17	1785	1.156	861	384	159	216	14	4	69	9	10	59
18	6699	4.340	3498	1264	570	833	42	8	161	37	47	239
19	3984	2.581	2008	797	372	479	17	12	90	41	32	136
20	870	0.564	364	164	92	153	20	14	12	6	22	23
21	1660	1.075	651	375	236	182	36	37	35	19	34	55
22	814	0.527	394	121	79	94	7	6	58	8	20	27
23	60	0.039	15	18	7	7	3	2	2	1	5	0
合計	154353	100.000	74855	30431	14676	19911	1798	900	4141	1236	1473	4932

(18.532%)，7(工学)23024 編(14.916%)で，上位 3 分野で全体の 52.362% を占めている．その他，5% を超える分野は，11(物質科学)，14(分子生物学および遺伝学)，15(総合)がある．また，国別に見た場合においても，3(化学)，4(臨床医学)，7(工学)の 3 研究分野は同様に強く出ている．

次に，抽出された論文に基づいて，研究者の専門分野を特定する．表 8 に国別の研究者の専門分野を示す．専門分野が特定できた著者は 4992 名であった．研究分野別では，1 位は，5(コンピューター科学)の 1811 名で 36.278% であった．以下，7(工学)1488 名(29.808%)，3(化学)818 名(16.386%)で，上位 3 分野で全体の 82.472% を占めている．その他，5% を超える分野は，

表 8. IoT 関連論文における著者の専門分野（2016）.

			1	2	3	4	5	6	7	8	9	10
研究分野	合計	比率[%]	中国	アメリカ	韓国	イギリス	イタリア	スペイン	日本	インド	ドイツ	台湾
1	22	0.441	12	1	2	0	1	0	1	4	0	1
2	38	0.761	12	6	8	3	1	3	1	2	1	1
3	818	16.386	239	138	107	69	58	121	25	13	23	25
4	394	7.893	155	81	44	26	19	10	28	9	12	10
5	1811	36.278	403	324	262	210	194	123	60	56	105	74
6	83	1.663	10	16	2	9	24	0	1	2	17	2
7	1488	29.808	359	270	144	117	120	121	117	74	101	65
8	19	0.381	4	8	3	3	0	1	0	0	0	0
9	38	0.761	14	15	2	0	2	3	1	0	1	0
10	0	0.000	0	0	0	0	0	0	0	0	0	0
11	25	0.501	4	11	1	2	0	0	4	1	2	0
12	15	0.300	3	2	0	0	9	0	0	0	1	0
13	1	0.020	1	0	0	0	0	0	0	0	0	0
14	39	0.781	9	4	4	1	2	2	4	13	0	0
15	34	0.681	6	7	3	5	2	0	4	3	2	2
16	6	0.120	0	1	0	0	1	0	2	0	1	1
17	1	0.020	0	0	1	0	0	0	0	0	0	0
18	44	0.881	16	6	2	6	3	1	0	0	0	10
19	8	0.160	1	0	0	1	1	1	1	1	2	0
20	3	0.060	0	0	2	0	0	0	0	0	1	0
21	85	1.703	17	25	8	14	8	9	0	2	1	1
22	5	0.100	0	0	1	1	0	0	1	1	1	0
23	15	0.300	0	3	0	4	4	0	0	0	4	0
合計	4992	100.000	1265	918	596	471	449	395	250	181	275	192
			上位 3 分野の割合 [%]									
	調和平均[%]		1	2	3	4	5	6	7	8	9	10
3	12.673		18.893	15.033	17.953	14.650	12.918	30.633	10.000	7.182	8.364	13.021
5	34.931		31.858	35.294	43.960	44.586	43.207	31.139	24.000	30.939	38.182	38.542
7	30.908		28.379	29.412	24.161	24.841	26.726	30.633	46.800	40.884	36.727	33.854

4(臨床医学)のみであった．また，国別に見た場合においても，上位 10 か国において 5(コンピューター科学)が 8 か国で 1 位，7(工学)が 10 か国で 2 位以上，3(化学)が 9 か国で 3 位以上となり上位を占めている．個々の特徴では，スペインは，3(化学)を専門とする著者が 2 位で 30% 台であるが，1 位と 1% 以内の差であり多い傾向がある．次に，韓国，イギリス，および，イタリアは，5(コンピューター科学)を専門とする著者が 1 位で 40% 台であり多い傾向がある．最後に，日本とインドは，7(工学)を専門とする著者が 1 位で 40% 台であり多い傾向がある．

5.3.3 研究分野間のつながり

A. 研究分野間つながりの見える化

最後のステップ❺では，上位 10 か国において，組織レベルの異分野間のつながりの分析を行う．本稿では，組織とは国または地域，そして，異分野間のつながりとは各研究分野間における研究者を介したつながりであり，分野間の知識共有として示している．また，各研究分野間のつながりにおいて，媒介する研究者の数が多い，または，比率が高い場合にそれら分野間の知識の共有が進むとの理解のもと接続線の幅を太く示している．なお，分野間のつながりは計 253 通りである．

次に IoT 論文数上位 10 か国における研究分野間のつながりの平均を示す．このつながりでは，各国の研究者数の違いを相殺するために各国内でのつながりの比率を用いている．図 6 に，IoT 論文数上位 10 か国における研究分野間のつながりの平均を示す．図 6 における線の太さは，各国の研究分野間の接続率の平均値を示している．図 6(a) はすべてのつながり，図 6(b)

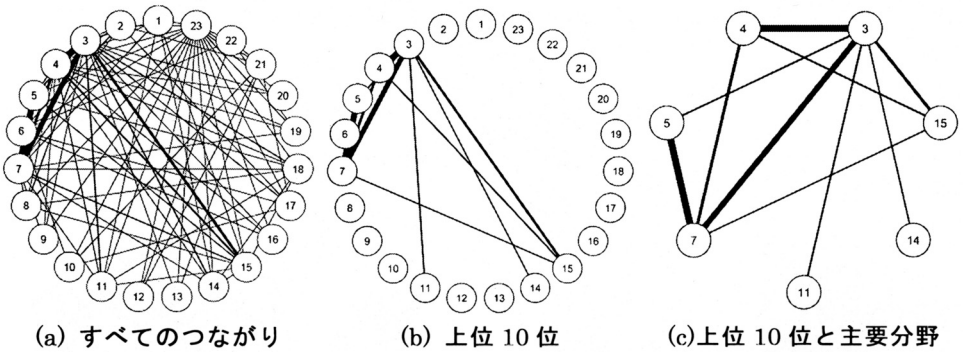


図 6. IoT 論文数上位 10 か国における研究分野間の平均的つながり（2016）.

は上位 10 位までのつながり，そして，図 6(c) は上位 10 位までのつながりと主要分野のみを示している．

図 6(c) に着目すると，化学と臨床医学と工学と総合間 [3-4-7-15]，化学とコンピューター科学と工学間 [3-5-7] が，すべての研究分野と連結する完備ネットワーク（complete network）であり強いつながりがあることを示している．そして，化学は物質科学間 [3-11] と分子生物学&遺伝学間 [3-14] と個別の強いつながりを示している．

B. 研究分野間つながりの相対比較と分析対象の選択

図 6 の表示手法は，研究分野間のつながりを見える化することが可能である．一方，分野間のつながりの主な特徴を比較して定量的に示すことが困難である．そこで，本稿では，上位 10 か国の分野間つながりに対して，分野間つながりを個体，国を変数として相関行列を用いた主成分分析を適用，得られた主成分得点の散布図により，重要な分野間つながりを抽出しその特徴を示す．

表 9 にすべての研究分野間のつながりの主成分分析における各主成分の寄与率を示す．各主成分の寄与率に着目すると，表 9 より，寄与率が 0.10 以上であること，さらに，累積寄与率が 0.852 であることから，本稿では第 2 主成分までを分析の対象とした．

図 7 に IoT 論文数上位 10 か国のすべての研究分野間つながりの主成分分析の結果を示す．図 7 では，すべての研究分野間つながりを示しているが，多くが原点付近に位置しており国別比較に寄与しないと考えられる．例えば，第 1 主成分と第 2 主成分において，どちらかの絶対値が 0.250 以上の研究分野間のつながりは 49，0.500 以上は 23，1.000 以上は 14 である．そこで本稿では，多くの分野間つながりは国別比較に寄与しないと考え，以後の分析は，研究分野間つながりにおける論文数の上位 10 か国における研究分野間つながりにおいて，各国の上位 10 位までの合計 36 の研究分野間つながりのみを対象とした．以後の分析は，断りがない場合，この 36 の研究分野間つながりを分析対象としている．

C. 選択された研究分野間つながりの相対比較

表 10 に IoT 論文数上位 10 か国に関する 36 の研究分野間のつながりの主成分分析における各主成分の寄与率と累積寄与率を示す．各主成分の寄与率に着目すると，寄与率が 0.10 以上であること，さらに，累積寄与率が 0.820 であることから，本稿では第 2 主成分までを分析の対象とした．

各主成分の名称を決めるために因子負荷量を示す．図 8 に横軸を第 1 主成分，縦軸を第 2 主成分として因子負荷量を示す．なお，因子負荷量は，各主成分と各変数のピアソン相関係数である．また，本稿では，因子負荷量の絶対値が 0.700 以上を変数に対する因子の影響が強いと

表 9. すべての研究分野間のつながりの主成分分析における寄与率（2016）.

主成分	1	2	3	4	5	6	7	8	9	10
寄与率	0.711	0.141	0.068	0.026	0.019	0.015	0.009	0.006	0.003	0.001
累積寄与率	0.711	0.852	0.920	0.946	0.966	0.981	0.990	0.996	0.999	1.000

表 10. 研究分野間のつながりの主成分分析における寄与率（2016）.

主成分	1	2	3	4	5	6	7	8	9	10
寄与率	0.635	0.185	0.090	0.027	0.025	0.017	0.010	0.006	0.004	0.001
累積寄与率	0.635	0.820	0.910	0.937	0.962	0.979	0.990	0.996	0.999	1.000

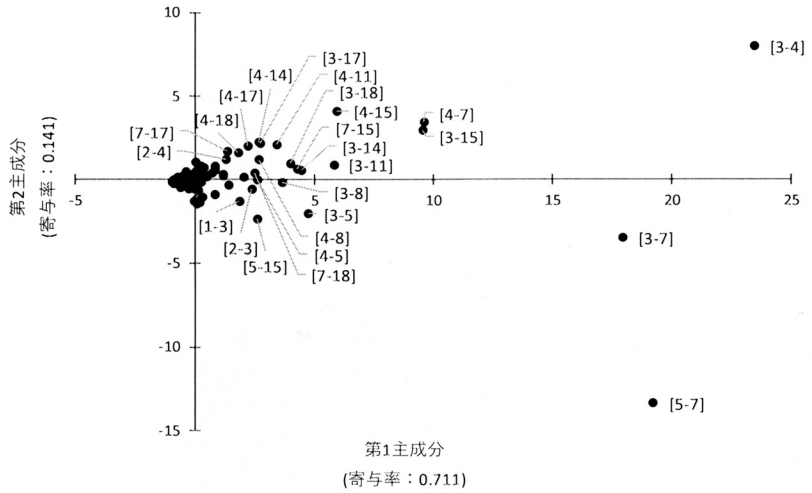


図 7. すべての研究分野間のつながりの主成分分析（2016）. 注意：253 通りの研究分野間のつながりが対象.

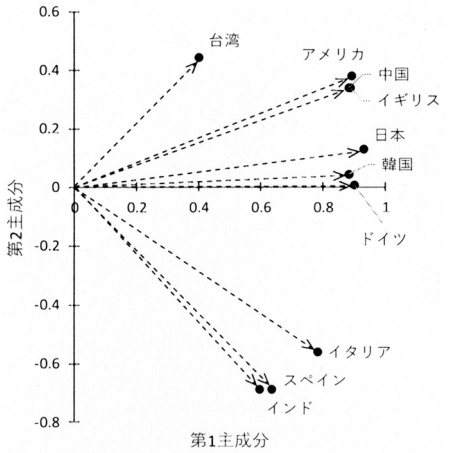
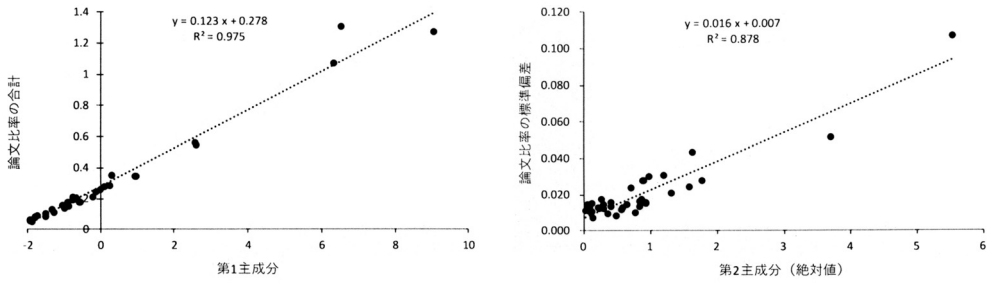


図 8. 研究分野間のつながりの主成分分析における因子負荷量（2016）.



(a)第 1 主成分得点　　(b)第 2 主成分得点

図 9．分野間のつながりの平均および標準偏差と主成分分析の関係（2016）．

して，0.300 以上を中程度，0.300 未満を弱いとする．各主成分の構成は因子負荷量の絶対値が 0.300 以上のもので構成すると，第 1 主成分は全ての国の因子負荷量が中程度以上，第 2 主成分は韓国(3 位)日本(7 位)ドイツ(9 位)の因子負荷量が弱いが，その他の国は中程度以上である．

第 1 主成分，第 2 主成分共に因子負荷量が中程度以上の国が多いため，名称を決めることが困難である．そこで，新たな情報として，研究分野間のつながりと主成分の関係に着目してその特徴を示す．図 9(a)に上位 10 か国の研究分野間のつながりの論文比率の合計と第 1 主成分の関係を示す．研究分野間のつながりの論文比率の合計と第 1 主成分の関係は，切片が 0.278，傾きが 0.123，決定係数が 0.975 であり，相関関係が強い結果を得た．続いて，図 9(b)に上位 10 か国の研究分野間のつながりの論文比率の標準偏差と第 2 主成分(絶対値)の関係を示す．研究分野間のつながりの論文比率の標準偏差と第 2 主成分の関係は，切片が 0.007，傾きが 0.016，決定係数が 0.878 であり，相関関係が強い結果を得た．ただし，第 2 主成分は絶対値を用いている．第 2 因子の主成分の符号は，研究分野間のつながりの論文比率における分布の歪度と一致しているが，歪度の情報をのぞいた比較を行った．

以上の因子負荷量の分析結果，そして，研究分野間つながりとの分析結果から，各主成分の特徴は，第 1 主成分は「分野間のつながりの強さ」，第 2 主成分は「分野間のつながりのバラツキ度」でまとめられる．

図 10 に IoT 論文数上位 10 か国に関する研究分野間のつながりの主成分分析の結果を示す．第 1 主成分において，15 以上で示されている分野間つながりの化学と臨床医学間 [3-4]，化学と工学間 [3-7]，および，コンピューター科学と工学間 [5-7] は，図 6 にて分野間のつながりが強いものとして強調表示されている．次に第 2 主成分において，±5 以上で示されている分野間つながりの化学と臨床医学間 [3-4]，および，コンピューター科学と工学間 [5-7] は，図 7 にて分野間のつながりが強いものとして強調表示されているが，上位 10 か国の研究分野間のつながりの標準偏差が，それぞれ 3.800，8.780 であり，他のつながりに比べて大きな値を示している．

分析の結果，分野間のつながりの強いものとして，化学と臨床医学 [3-4]，コンピューター科学と工学間 [5-7]，および，化学と工学間 [3-7] が抽出された．次に中程度の分野間のつながりがあるものとして，臨床医学と工学間 [4-7]，化学と総合間 [3-15] ，臨床医学と総合間 [4-15]，化学と物質科学間 [3-11] 等が抽出された．このように国別で分野間のつながりに差があることが示唆された．

5.3.4　研究分野間のつながりの国別比較

図 11 に IoT 論文数上位 10 か国の研究分野間のつながりを示す．なお，図 11 は，各国上位 10 位までのつながりと**主要分野**のみ表示している．図 11(a)の中国は，化学と臨床医学と工学

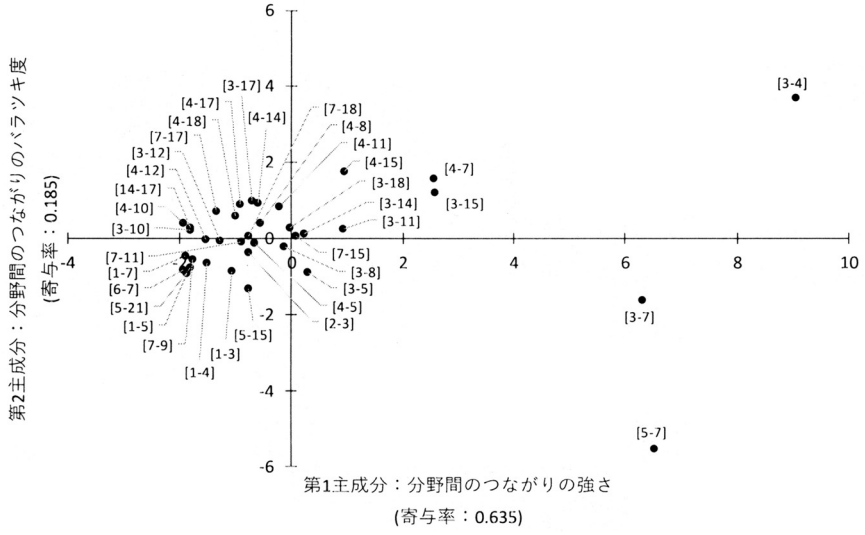


図 10. 研究分野間のつながりの主成分分析（2016）.

間 [3-4-7]，化学と臨床医学と総合間 [3-4-15]，化学と臨床医学と物理学間 [3-4-18] が，すべての研究分野と連結する完備ネットワーク（complete network）である．また，化学と物質科学間 [3-11] と化学と分子生物学&遺伝学間 [3-14]，コンピューター科学と工学間 [5-7] につながりがある．図 11（b）のアメリカは，化学と臨床医学と工学と総合間 [3-4-7-15]，化学と臨床医学と物質科学間 [3-4-11] が完備ネットワークである．また，臨床医学と分子生物学&遺伝学間 [4-14]，コンピューター科学と工学間 [5-7] につながりがある．図 11（c）の韓国は，化学と臨床医学と工学間 [3-4-7]，化学と臨床医学と環境/生態学間 [3-4-8]，化学と臨床医学と物質科学間 [3-4-11]，化学と臨床医学と薬理学&毒物学間 [3-4-17] が完備ネットワークである．また，コンピューター科学と工学間 [5-7] につながりがある．

図 11（d）のイギリスは，化学と臨床医学と工学と総合間 [3-4-7-15]，化学と臨床医学と物理学間 [3-4-18] が完備ネットワークである．また，臨床医学は物質科学間 [4-11] と分子生物学&遺伝学間 [4-14]，コンピューター科学と工学間 [5-7] につながりがある．図 11（e）のイタリアは，化学と臨床医学とコンピューター科学間 [3-4-5]，化学とコンピューター科学と工学間 [3-5-7]，化学とコンピューター科学と総合間 [3-5-15] が完備ネットワークである．また，環境/生態学間 [3-8]，物質科学間 [3-11]，分子生物学&遺伝学間 [3-14] につながりがある．図 11（f）のスペインは，化学と臨床医学と工学間 [3-4-7]，化学とコンピューター科学と工学間 [3-5-7]，化学と工学と環境/生態学間 [3-7-8] が完備ネットワークである．また，化学と生物学&生化学間 [2-3]，コンピューター科学と社会科学・一般間 [5-21]，工学は地球科学間 [7-9] と物理学間 [7-18] につながりがある．

図 11（g）の日本は，化学と臨床医学と工学間 [3-4-7]，化学と臨床医学と数学間 [3-4-12] が完備ネットワークである．また，化学は環境/生態学間 [3-8]，物質科学間 [3-11]，分子生物学&遺伝学間 [3-14]，総合間 [3-15]，コンピューター科学と工学間 [5-7] につながりがある．図 11（h）のインドは，農学と化学と臨床医学間 [1-3-4]，農学と臨床医学とコンピューター科学と工学間 [1-4-5-7]，コンピューター科学と工学と総合間 [5-7-15] が完備ネットワークである．また，化学は工学間 [3-7] と総合間 [3-15]，工学は経済学&ビジネス間 [7-6] と物質科学間 [7-11] につながりがある．図 11（i）のドイツは，化学とコンピューター科学と工学間 [3-5-7]，化学と工学と物理学

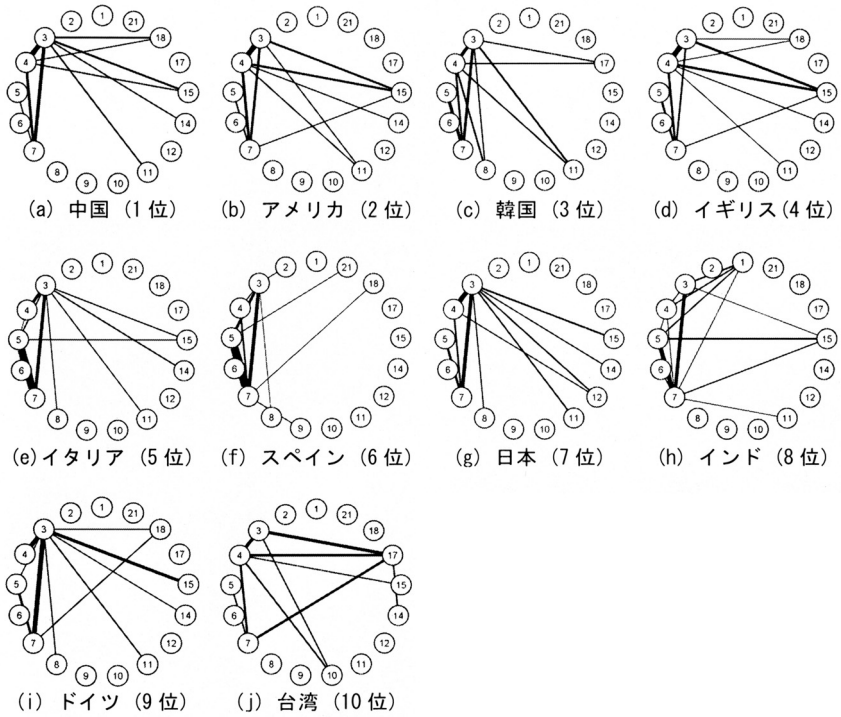


図 11. IoT 論文数上位 10 か国の研究分野間のつながりと主要分野 (2016)．注意 1：表示されている研究分野は，いずれかの国の関連研究分野である．注意 2：10 位の研究分野間つながりが複数ある場合，10 本以上のつながりで示す．イギリスは 11 本，インドは 14 本である．

間 [3-7-18] が完備ネットワークである．また，化学は臨床医学間 [3-4]，環境/生態学間 [3-8]，物質科学間 [3-11]，分子生物学&遺伝学 [3-14]，総合 [3-15] につながりがある．図 11(j) の台湾は，化学，臨床医学，免疫学 [3-4-10]，化学，臨床医学，薬理学&毒物学間 [3-4-17]，化学，工学，薬理学&毒物学間 [4-7-17] が完備ネットワークである．また，臨床医学と総合間 [4-15]，コンピューター科学と工学間 [5-7]，分子生物学&遺伝学と薬理学&毒物学間 [14-17] につながりがある．

A. 多様性の国別比較

本稿では，研究分野間のつながりにおいて，少数のつながりに集中している国を多様性が低い国，複数のつながりに分散している国を多様性が高い国とする．また，多様性の指標は，ローレンツカーブ (Lorenz, 1905) から導き出されたジニ係数 (Gini, 1936) を用いる．

ジニ係数は，社会科学分野において富の集中度の分析に用いられるが，Mizukami et al. (2017) は，研究 IR 分野における研究力測定において適用している．図 12 に IoT 論文数上位 10 か国の研究分野間のつながりの多様性を示す．図 12(a) はローレンツカーブ，図 12(b) はローレンツカーブから導き出されたジニ係数を示す．

分析の結果，研究分野間のつながりの多様性は相対的に 3 グループに大別できる．まず多様性が高いグループとして，ジニ係数が 0.520 未満の韓国が挙げられる．次に多様性が中程度のグループとして，ジニ係数が 0.650 未満の中国，アメリカ，イギリス，日本，インド，台湾，が挙げられる．最後に多様性が低いグループとして，ジニ係数が 0.650 以上のイタリア，スペイン，ドイツが挙げられる．なお，これら閾値は，国別の傾向を明確に分類できる値としている．

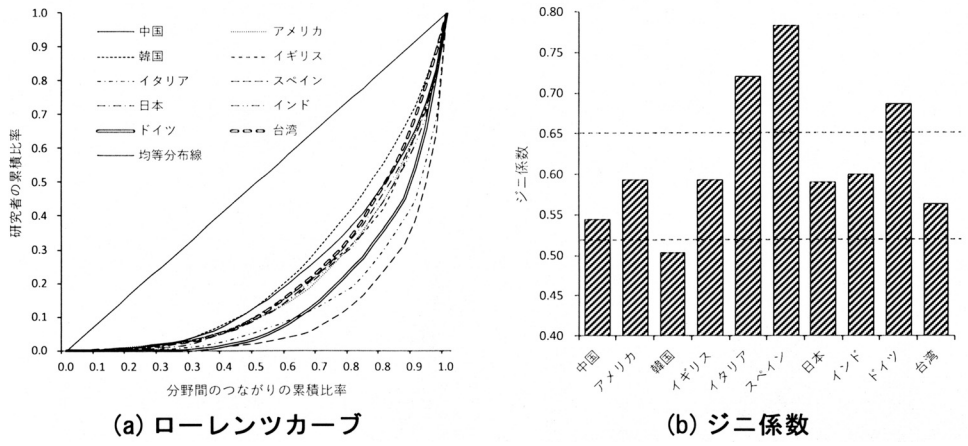


図 12. IoT 論文数上位 10 か国の研究分野間のつながりの多様性 (2016).

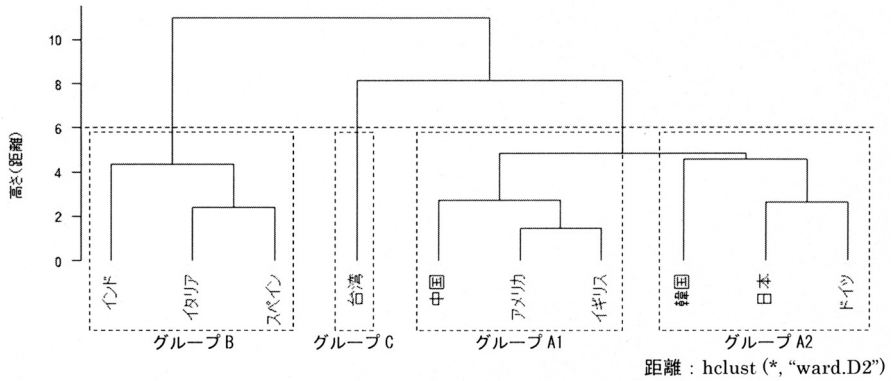


図 13. 研究分野間のつながりに関する IoT 論文数上位 10 か国のデンドログラム.

B. 類似性の国別比較

本稿では，上位 10 か国それぞれ上位 10 種類の分野間つながりに対して，国を個体，分野間つながりを変数として階層型クラスター分析を施して，デンドログラムで可視化する．図 13 に研究分野間のつながりに関する IoT 論文数上位 10 か国のデンドログラムを示す．なお，クラスター分析におけるクラスター間距離はウォード法を用いた．

図 13 のデンドログラムにて，IoT 上位 10 か国を 3 つのグループに分類した．まず，中国，アメリカ，韓国，イギリス，日本，ドイツのグループ A，そして，インド，イタリア，スペインのグループ B，最後に台湾単独のグループ C である．また，グループ A には 2 つのクラスターがあり，中国，アメリカ，イギリスのサブグループ A1，韓国，日本，ドイツのサブグループ A2 に分割できる．なお，サブグループ A2 において，韓国の距離が遠いことが示されており，日本とドイツ間の類似性に比べて，韓国との類似性が低いからであると考えられる．

しかし，図 13 の階層型クラスター分析のデンドログラムでは，グループ化はできるが，その分類の要因を把握することができない．そこで，本稿では，上位 10 か国の分野間つながりに対して，国を個体，分野間つながりを変数として分散共分散行列を用いた主成分分析を適用して国別の散布図を示し分類の要因の抽出とその特徴を示す．表 11 に IoT 論文数上位 10 か国

表 11．研究分野間のつながりの主成分分析における寄与率（2016）．

主成分	1	2	3	4	5	6	7	8	9	10
寄与率	0.476	0.251	0.089	0.065	0.054	0.031	0.020	0.011	0.003	0.000
累積寄与率	0.476	0.727	0.816	0.881	0.935	0.966	0.986	0.997	1.000	1.000

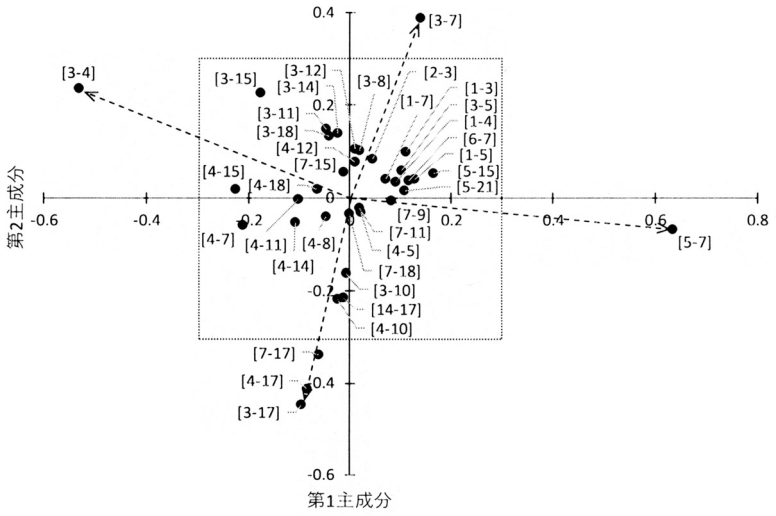


図 14．研究分野間のつながりの主成分分析における因子負荷量．注意：点線の四角は ±0.3 の範囲を示している．

に関する国別の主成分分析における各主成分の寄与率と分野間つながりの因子負荷量を示す．各主成分の寄与率に着目すると，表 11 より，寄与率が 0.10 以上であること，さらに，累積寄与率が 0.727 であることから，本稿では第 2 主成分までを分析の対象とした．寄与率の合計は 0.727 である．

各主成分にその特徴を示す名称を付ける目的にて，各主成分と因子負荷量および研究分野間つながりとの関係を示す．まず，図 14 に横軸を第 1 主成分，縦軸を第 2 主成分として因子負荷量を示す．また，本稿では，因子負荷量の絶対値が 0.700 以上を変数に対する因子の影響が強いとして，0.300 以上を中程度，0.300 未満を弱いとする．

各主成分の構成は因子負荷量の絶対値 0.300 以上のもので構成すると，第 1 主成分は，化学と臨床医学間 [3-4]，コンピューター科学と工学間 [5-7] を合成したものである．第 2 主成分は，薬理学&毒物学と化学間 [3-17]，臨床医学間 [4-17]，工学間 [7-17]，そして，化学と工学間 [3-7] を合成したものである．

次に，図 15 に研究分野間のつながりに関する IoT 論文数上位 10 か国の主成分分析の結果を示す．本稿では，上位 10 か国の分野間つながりに対して，国を個体，分野間つながりを変数として分散共分散行列を用いた主成分分析を適用する．主成分分析の結果，図 15 の国々のグループは，図 13 のデンドログラムのグループと同様の結果であり，さらに，その位置関係からグループ化の要因を示している．

5.4　まとめ

表 12 に 5 章での分析結果として，IoT 論文数上位 10 か国のグループ化とその要因をまとめたものを示す．本稿では，多様性として，少数の分野間つながりに特化した国，多くの分野間

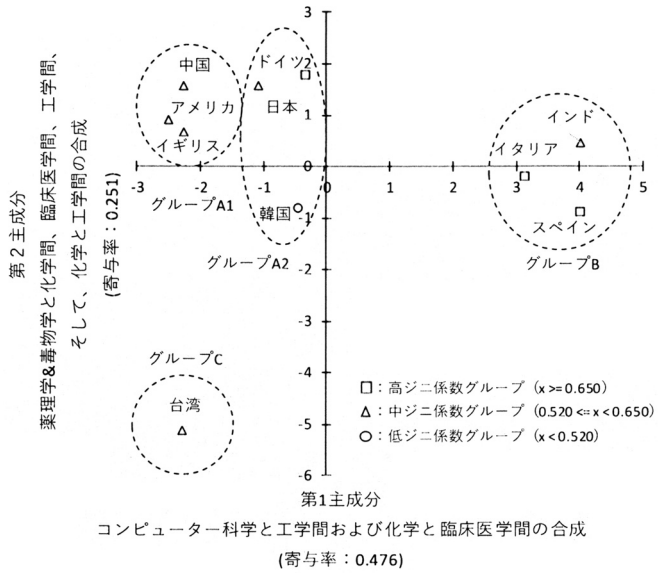


図 15. 研究分野間のつながりに関する IoT 論文数上位 10 か国の主成分分析（2016）.

表 12. IoT 論文数上位 10 か国のグループ化とその要因．研究分野：[3] 化学，[4] 臨床医学，[5] コンピューター科学，[7] 工学，[17] 薬理学&毒物学．注意：背景が灰色のデータは 0.100 以上の比率である．#1：全研究分野間つながり 253 組における比率である．

Gr.	国	論文数	ジニ	主成分得点		国別の研究分野間つながり比率#1					
	地域	順位	係数	第１主成分	第２主成分	[3-4]	[3-7]	[3-17]	[4-17]	[5-7]	[7-17]
A1	中国	1	0.544	-2.270	1.579	0.113	0.075	0.005	0.003	0.025	0.002
A1	アメリカ	2	0.593	-2.507	0.921	0.137	0.062	0.004	0.006	0.042	0.002
A1	イギリス	4	0.594	-2.255	0.661	0.140	0.045	0.006	0.007	0.056	0.002
A2	韓国	3	0.504	-0.446	-0.787	0.068	0.055	0.028	0.030	0.063	0.017
A2	日本	7	0.591	-1.083	1.561	0.123	0.091	0.010	0.000	0.057	0.009
A2	ドイツ	9	0.600	-0.345	1.785	0.124	0.134	0.005	0.003	0.066	0.000
B	イタリア	5	0.720	3.145	-0.186	0.097	0.105	0.008	0.000	0.257	0.000
B	スペイン	6	0.783	4.024	-0.873	0.032	0.117	0.000	0.005	0.276	0.000
B	インド	8	0.600	4.022	0.443	0.035	0.112	0.005	0.002	0.139	0.002
C	台湾	10	0.564	-2.284	-5.104	0.075	0.017	0.072	0.063	0.033	0.055

つながりに分散した国を把握するためにジニ係数を用い，ジニ係数が 0.520 未満である国を多様性が相対的に高い国，0.650 未満である国を多様性が相対的に中程度の国，そして，0.650 以上である国を多様性が相対的に低い国とした．なお，これら閾値は，国別の傾向を明確に分類できる値としている．分析の結果，IoT 論文数上位 10 か国において，多様性が高い国として中国，韓国，台湾の 3 つのアジア諸国，多様性が中程度の国として，アメリカ，イギリス，日本，インド，ドイツの 5 か国，多様性が低い国として，イタリア，スペインの 2 つの EU 諸国とした．

次に，主な分野間つながりの傾向により IoT 論文数上位 10 か国を 3 つのグループに分類した．まず，グループ A は，経済が発展している 1 位の中国，2 位のアメリカ，3 位の韓国，4 位のイギリス，7 位の日本，9 位のドイツのグループである．特徴は，化学と臨床医学間 [3-4] の比率が高く，工学とコンピューター科学間 [5-7] の比率が低くなる傾向にある．そして，サブグループ A1 はグループの特徴が強く，サブグループ A2 は弱いグループである．なお，サブ

グループ A2 の韓国は，日本とドイツに比べてグループの傾向がさらに弱くなっている．これは，図 13 の階層型クラスター分析の結果から，サブグループ A2 において，韓国の距離が遠いことが示されており，この違いにより韓国の傾向が異なっていると考えられる．続くグループ B は，南ヨーロッパの 5 位のイタリア，6 位のスペイン，8 位のインドのグループである．特徴は，グループ A とは反対に化学と臨床医学間 [3-4] の比率が低く，工学とコンピューター科学間 [5-7] の比率が高くなる傾向にある．最後のグループ C は，アジアの新興国である 10 位の台湾単独のグループである．特徴は，薬理学&毒物学と化学間 [3-17]，臨床医学間 [4-17]，工学間 [7-17] の比率が高く，化学と工学間 [3-7] の比率が低く出る傾向にある．

6. 考察

2011 年，ドイツ技術科学アカデミーとドイツ連邦教育科学省は，「あらゆる社会システムの効率化」「新産業の創出」「知的生産性の向上」を目指した技術的フレームワーク Industry 4.0 を発表した．本稿は，Industry 4.0 において，フィジカル空間とサイバー空間の橋渡しをする IoT 技術の研究に着目，異分野融合による研究促進の視点から，各国の研究促進としてのイノベーションの戦略を分析して考察するものである．

分析では，一連の研究で導出された分析手法と共に，主成分分析，階層型クラスター分析を行い，多角的な分析と考察を展開するものである．なお，現状の把握という目的では，直近の年のデータを分析するのが望ましいが，書誌データベースの特性上，直近数年は，データが更新されていないなど，ノイズが多い場合があるため，データが安定していると考えられる 2016 年を対象年とした．

分析の結果，IoT 研究の異分野融合では，分野間つながりの類似性から，「工学とコンピューター科学の連携の強さ」という技術的アプローチ，「化学と臨床医学の連携の強さ」というアプリケーション的アプローチ，「薬理学&毒物学を介した化学，臨床医学，工学の連携の強さ」という薬理学&毒物学を中心としたアプローチにより国を次の 3 つのグループに分類した．まず，アプリケーション的アプローチの中国，アメリカ，韓国，イギリス，日本，ドイツの 6 か国のグループ，技術的アプローチのイタリア，スペイン，インドの 3 か国のグループ，そして，薬理学&毒物学を中心としたアプローチとアプリケーション的アプローチの台湾単独のグループである．また，ドイツ，イタリア，スペインの EU3 か国は一部の分野間の研究に集中，中国と韓国のアジア 2 か国は広く分散させた研究を展開していることを示した．

日本は，分野間研究の量では 7 位，質ではアプリケーション的アプローチ，多様性では中程度に分散させた研究を展開していることが判明した．また，日本と同様の質と多様性を持つ国として 2 位のアメリカと 4 位のイギリス，そして，日本と同様の質であるが，一部の分野間の研究に集中している国として 8 位のインドと 9 位のドイツが挙げられる．

以上のように，本稿で提案した手法は，本稿は，IoT 研究を推進するための異分野融合によるイノベーション戦略において，量，質，そして集中度という指標を提示して，その発展の推進に貢献するものであると考えられる．今後の研究の方向性は 2 つある．まず，国際比較と組織間の分析事例を増やし，それら手法の効果を検討することがある．そして，これら手法の確立と共にその普及に努めることがある．

謝　辞

一連の研究に協力いただいた統計数理研究所の金藤浩司先生，北村浩三先生，本多啓介先生，岡本基先生，工学院大学の鈴木重徳先生に感謝する．また，原稿を注意深くお読み頂き適切な助言を頂いたことに対して，二人の匿名査読者および編集委員に感謝する．本研究は JSPS 科

研費 JP17K04710，統計数理研究所共同研究プログラム(2019-ISMCRP-1026)の助成を受けたものである．本研究は，日本大学 生産工学部異分野融合イノベーションリサーチ・グループのプロジェクトである．

参 考 文 献

藤垣裕子, 平川秀幸, 富澤宏之, 調麻佐志, 林隆之, 牧野淳一郎 (2004).『研究評価・科学論のための科学計量学入門』, 丸善, 東京.

Gini, C. (1936). On the measure of concentration with special reference to income and statistics, *Colorado College Publication*, **208**, 73–79.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth, *Publications of the American Statistical Association*, **9**(70), 209–219.

Mizukami, Y., Honda, K., Suzuki, S., Nakano, J. and Otabe, A. (2016). Co-author information and authors' affiliation information in scientific literature using centralities—The researchers who act as mediators between organizations—, *International Journal of the Japan Association for Management Systems*, **8**(1), 1–8.

Mizukami, Y., Mizutani, Y., Honda, K., Suzuki, S. and Nakano, J. (2017). An international research comparative study of the degree of cooperation between disciplines within mathematics and mathematical sciences: Proposal and application of new indices for identifying the specialized field of researchers, *Behaviormetrika*, **44**, 385–403.

Mizukami, Y., Honda, K. and Nakano, J. (2018). Study on research trends on the internet of things using network analysis, *International Journal of the Japan Association for Management Systems*, **10**(1), 27–35.

文部科学省 (2019). 研究力向上改革 2019, http://www.mext.go.jp/a_menu/other/1416069.htm(最終アクセス 2020/5/1).

根岸正光, 山崎茂明 (2001).『研究評価—研究者・研究機関・大学におけるガイドライン』, 丸善, 東京.

Porter, A. L. and Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time, *Scientometrics*, **81**, 719–745.

Putnik, G. D., Ferreira, L., Lopes, N. and Putnik, Z. (2019). What is a cyber-physical system: Definitions and models spectrum, *FME Transactions*, **47**(4), 663–674.

Rafols, I. and Meyer, M. (2010). Diversity measures and network centralities as indicators of interdisciplinarity: Case studies in bionanoscience, *Scientometrics*, **82**, 263–287.

Stirling, A. (2007). A general framework for analyzing diversity in science, technology and society, *Journal of the Royal Society Interface*, **4**, 707–719.

Vergidis, P., Karavasiou, A., Paraschakis, K., Bliziotis, I. and Falagas, M. (2005). Bibliometric analysis of global trends for research productivity in microbiology, *European Journal of Clinical Microbiology and Infectious Diseases*, **24**(5), 342–346.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J. and Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature, *Journal of Informetrics*, **5**(1), 14–26.

Using an Academic Literature Database to Evaluate International Interdisciplinary Fusion in IoT Research through Coauthor Analysis

Yuji Mizukami[1] and Junji Nakano[2,3]

[1]College of Industrial Technology, Nihon University
[2]Faculty of Global Management, Chuo University
[3]The Institute of Statistical Mathematics

In 2011, the German Academy of Technology and the German Federal Ministry of Education and Science announced the Industry 4.0 technical framework, which aims to make all social systems more efficient, create new industries, and improve intellectual productivity. The foundational technologies of Industry 4.0 are the Internet of Things (IoT), big data, and artificial intelligence. This paper focuses on research on IoT technology that bridges the physical space and cyberspace, and analyzes each country's research promotion strategy from the standpoint of integrating different fields. In our analysis, we conducted an international comparison to examine the level of fusion among different domains of IoT research. We considered varied perspectives and approaches, and employed methods derived from a series of studies to perform principal component analysis and hierarchical clustering analysis.

Key words: IR, research ability, coauthor analysis, innovation.

統計数理(2020)
第 68 巻 第 2 号 287–303

［原著論文］

グループ正則化に基づく順序ロジットモデルにおける隣接クラスの統合

永沼 瑞穂[1]・吉川 剛平[2]・川野 秀一[2]

(受付 2020 年 3 月 23 日；改訂 6 月 29 日；採択 6 月 29 日)

要　　旨

本稿では，多クラス分類を目的とした順序ロジットモデルの枠組みにおいて，クラスを統合する方法を提案する．クラスを統合することによって，モデルの解釈が容易になるとともに，冗長なクラスが存在する場合はその冗長性を排除することができる．クラスの統合は隣接する各クラスの事後確率が等しくなるときに実行し，この目的のために構造的正則化の一つであるグループ正則化を用いる．グループ正則化は微分不可能な点を含むため，モデルに含まれるパラメータの推定値は，交互方向乗数法に基づく推定アルゴリズムにより得る．正則化パラメータの値は，グループ正則化により推定されたモデルの自由度をもつベイズ型情報量規準により選択する．モンテカルロ・シミュレーションおよび実データへの適用を通して，提案手法の有効性を検証する．

キーワード：グループ lasso，交互方向乗数法，順序付きカテゴリカルデータ，隣接カテゴリーロジットモデル．

1. はじめに

多クラス分類は，自然科学・社会科学のさまざまな分野で用いられており，それを実行するための統計モデルは，クラスを表すカテゴリカルデータを目的変数，それに関係する共変量を説明変数としてモデル化される．多クラス分類においては，クラスの個数が多くなるときには注意が必要である．得られた解析結果の解釈が困難になる(Price et al., 2019)，または冗長なクラスが生じる，といった現象が起きやすくなる．その例として，マーケティング分野における顧客市場のセグメンテーションを挙げることができる．顧客市場のセグメンテーションでは，大規模データ分析を行うことで顧客市場が過剰に細分化されてしまい，クラス毎の解釈が困難になるという現象が生じる．これはユーザーの分析コストを増大させ，マーケティング資源を浪費してしまうという恐れがある．このようなクラスの個数が多い状況においては，クラスを統合することによって，分析コストを削減することができ有用である．近年，そのクラスの統合を目的として，Price et al. (2019)は，多項ロジットモデルと構造的正則化に基づいた統計手法 group fused multinomial logistic regression(GFMR)を提案している．

一方，カテゴリカルデータの中でも，順序関係を持つデータは順序付きカテゴリカルデータ

[1] 電気通信大学大学院 情報理工学研究科：〒 182–8585 東京都調布市調布ヶ丘 1–5–1（現 株式会社マクロミル：〒 108–0075 東京都港区港南 2–16–1 品川イーストワンタワー 11F）

[2] 電気通信大学大学院 情報理工学研究科：〒 182–8585 東京都調布市調布ヶ丘 1–5–1

と呼ばれている．例えば，企業の格付けや 5 段階で表される学生の成績，アンケートの選好度などがある．このような順序付きカテゴリカルデータと共変量との間をモデリングする統計手法として，順序ロジットモデル（Agresti, 2010）が広く用いられている．前述したクラス過多な状況から生じる問題は，順序付きカテゴリカルデータを目的変数としたときも同様な理由で生じる．したがって，この場合でもクラスを統合することは有用である．Price et al. (2019) では，順序付きカテゴリカルデータを目的変数に用いる場合でも GFMR によってクラス統合を行っている．しかし，GFMR は多項ロジットモデルに基づいており，順序付きカテゴリカルデータを目的変数とするならば，順序ロジットモデルを用いる方が望ましい．

本稿では，順序ロジットモデルの枠組みの下，クラスを統合するための統計手法を提案する．順序ロジットモデルの中でも，隣接カテゴリーロジットモデル（Goodman, 1984）を用い，クラスの統合は，隣接する各クラスの事後確率が等しくなるときに実行する．この目的のために，構造的正則化の一つであるグループ正則化，特にグループ lasso（Yuan and Lin, 2006）を用いる．グループ lasso の正則化項は微分不可能な点を含むため，モデルに含まれるパラメータの推定値は，交互方向乗数法（Boyd et al., 2011）に基づく推定アルゴリズムにより得る．構築したモデルには正則化パラメータが含まれるが，この値をグループ正則化により推定されたモデルの自由度をもつベイズ型情報量規準（Schwarz, 1978）により選択する．

本稿の構成は次の通りである．2 節では順序ロジットモデル，特に隣接カテゴリーロジットモデルを紹介する．3 節では提案手法とその推定アルゴリズム，正則化パラメータの選択方法，関連研究の GFMR について述べる．4 節ではモンテカルロ・シミュレーションおよび実データへの適用を通して，提案手法の有効性を検証する．5 節では本稿のまとめを行い，今後の課題について検討する．

2. 順序ロジットモデル

順序ロジットモデルは，各クラスの事後確率の対数オッズを，説明変数の線形結合で表すことによってモデル化される（Ananth and Kleinbaum, 1997; Agresti, 2010）．この対数オッズの設定方法によって，累積ロジットモデル（McCullagh, 1980; Armstrong and Sloan, 1989; Whitehead et al., 2001），連続比ロジットモデル（Fienberg and Mason, 1979; Fienberg, 2007），隣接カテゴリーロジットモデル（Adjacent-Categories Logit model, ACL）（Anderson, 1984; Goodman, 1984; Agresti, 1992）などのさまざまなモデルが提案されている．本稿ではこの中でも，隣接カテゴリーロジットモデルを用いる．

順序付きカテゴリカル変数 $G \in \{1, \ldots, J\}$ と p 次元説明変数 $\boldsymbol{x}^\dagger = (x_1, \ldots, x_p)^\top$ より，n 組のデータ $\{(g_i, \boldsymbol{x}_i^\dagger); i = 1, \ldots, n\}$ が得られたとする．ここで，順序付きカテゴリカル変数とは，何らかの尺度についてのクラスの順序を示したものである．例えば，アンケート選好度を変数とした場合，その取り得る選択肢として，1：あてはまる，2：ややあてはまる，3：ややあてはまらない，4：あてはまらない，の 4 つのクラスは，順序付きカテゴリカル変数である．また，説明変数に関するデータは標準化されているとする．つまり，$\sum_{i=1}^n x_{ij}^\dagger = 0$, $\sum_{i=1}^n (x_{ij}^\dagger)^2 = n$ $(j = 1, \ldots, p)$ が成り立つものとする．

説明変数に関するデータ $\boldsymbol{x}_i^\dagger$ が与えられた下で，j 番目のクラスに属する事後確率を $\Pr(g_i = j | \boldsymbol{x}_i^\dagger) = \pi_j(\boldsymbol{x}_i^\dagger)$ $(j = 1, \ldots, J)$ とする．このとき，隣接カテゴリーロジットモデルは

$$\log \frac{\pi_j(\boldsymbol{x}_i^\dagger)}{\pi_{j+1}(\boldsymbol{x}_i^\dagger)} = \beta_{0j} + \beta_{1j} x_{i1} + \cdots + \beta_{pj} x_{ip} = \boldsymbol{\beta}_j^\top \boldsymbol{x}_i, \qquad j = 1, \ldots, J-1 \tag{2.1}$$

となる．ここで，$\boldsymbol{x}_i = (1, \boldsymbol{x}_i^{\dagger\top})^\top$，$\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{pj})^\top$ は j 番目のクラスに対する $(p+1)$

次元の回帰係数ベクトルである．つまり，j 番目のクラスに属する確率と $(j+1)$ 番目のクラスに属する事後確率の対数オッズと，説明変数間をモデル化したものが隣接カテゴリーロジットモデルである．

いま，i 番目のデータ $\boldsymbol{x}_i^\dagger$ が j 番目のクラスに属しているとする．このとき，指示変数 y_{ij} $(i=1,\ldots,n;\ j=1,\ldots,J)$ を

$$y_{ij}=\begin{cases}1 & \text{データ } i \text{ がクラス } j \text{ に属しているとき}\\ 0 & \text{その他}\end{cases} \tag{2.2}$$

と定義すると，データ $\boldsymbol{x}_i^\dagger$ に対する順序付きカテゴリカル変数 G_i は，J 次元のベクトル $\boldsymbol{y}_i=(y_{i1},\ldots,y_{iJ})^\top$ と表すことができる．ここで，$y_{iJ}=1-\sum_{j=1}^{J-1}y_{ij}$ が成り立つことに注意しておく．

ベクトル $\boldsymbol{y}_i$ は(2.2)式の構成法から，確率 $\pi_1(\boldsymbol{x}_i^\dagger),\ldots,\pi_J(\boldsymbol{x}_i^\dagger)$ を持つ多項分布にしたがうことがわかる．したがって，隣接カテゴリーロジットモデルの対数尤度関数は，

$$\ell(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{J-1})=\sum_{i=1}^{n}\left\{\sum_{j=1}^{J-1}y_{ij}\sum_{k=j}^{J-1}\boldsymbol{\beta}_k^\top\boldsymbol{x}_i-\log\left[1+\sum_{j=1}^{J-1}\exp\left(\sum_{k=j}^{J-1}\boldsymbol{\beta}_k^\top\boldsymbol{x}_i\right)\right]\right\} \tag{2.3}$$

と表すことができる．(2.3)式を最大にするパラメータは，解析的に陽に表現することが難しいので，非線形最適化手法，例えば，ニュートン・ラフソン法によって求める．なお，(2.3)式の導出過程については Agresti (2010)を参照されたい．

3. 提案手法

本節では，隣接カテゴリーロジットモデルの枠組みの下，クラスを統合するために，モデルに含まれるパラメータを構造的正則化の一つであるグループ lasso (Yuan and Lin, 2006)に基づいて推定する方法を提案する．その上で，推定アルゴリズムならびに正則化パラメータの選択方法について述べる．また，関連研究として Price et al. (2019)の GFMR を紹介する．

3.1 クラス統合とグループ正則化対数尤度関数

Price et al. (2019)は，各クラスの事後確率が等しくなるときにクラスを統合することを提案した．本稿でもこの考えにしたがうものとする．しかし，本稿で扱っている隣接カテゴリーロジットモデルには，クラスに順序が存在しているため，隣接していないクラスを統合することはせず，隣接しているクラスのみを統合する．すなわち，本稿では，隣接した各クラスに対して，それらの事後確率が等しくなる場合にクラスを統合することを考える．以降，計画行列 $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^\top$ の階数は $(p+1)$，すなわちフルランクであると仮定する．

隣接カテゴリーロジットモデルにおいて，隣接したクラスの統合と回帰係数の関係について調べる．いま，j 番目のクラスと $(j+1)$ 番目のクラスを統合することを考える．すなわち，$\pi_j(\boldsymbol{x}_i^\dagger)=\pi_{j+1}(\boldsymbol{x}_i^\dagger)$ $(i=1,\ldots,n)$ となる状況を考える．このとき，

$$\log\frac{\pi_j(\boldsymbol{x}_i^\dagger)}{\pi_{j+1}(\boldsymbol{x}_i^\dagger)}=0,\qquad i=1,\ldots,n,$$

となる．この式と(2.1)式より，$\boldsymbol{\beta}_j^\top\boldsymbol{x}_i=0$ $(i=1,\ldots,n)$ となり，計画行列がフルランクであるという仮定を用いると $\boldsymbol{\beta}_j=\boldsymbol{0}$ となることがわかる．したがって，隣接カテゴリーロジットモデルにおいては，$\pi_j(\boldsymbol{x}_i^\dagger)=\pi_{j+1}(\boldsymbol{x}_i^\dagger)$ $(i=1,\ldots,n)$ ならば $\boldsymbol{\beta}_j=\boldsymbol{0}$ であることがわかる．一方，上述の手順を逆に考えることにより，$\boldsymbol{\beta}_j=\boldsymbol{0}$ ならば $\pi_j(\boldsymbol{x}_i^\dagger)=\pi_{j+1}(\boldsymbol{x}_i^\dagger)$ $(i=1,\ldots,n)$ というこ

とがわかり，以上より，$\pi_j(\boldsymbol{x}_i^\dagger) = \pi_{j+1}(\boldsymbol{x}_i^\dagger)$ $(i = 1, \ldots, n)$ となるための必要十分条件は $\boldsymbol{\beta}_j = \mathbf{0}$ である．なお，必要条件のみならば，計画行列のフルランク性の仮定は必要ないことに注意されたい．

したがってクラス統合を行うために，$\boldsymbol{\beta}_j = \mathbf{0}$ と推定することが可能なグループ正則化項が付与された対数尤度関数

$$\max_{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{J-1}} \left\{ \ell(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{J-1}) - \lambda \sum_{j=1}^{J-1} \|\boldsymbol{\beta}_j\|_2 \right\} \tag{3.1}$$

を最大化し，パラメータを推定する方法を提案する．ここで，$\ell(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{J-1})$ は(2.3)式の隣接カテゴリーロジットモデルに対する対数尤度関数，λ は非負の値をとる正則化パラメータ，$\|\cdot\|_2$ はベクトルの L_2 ノルムである．(3.1)式の第 2 項目は，隣接したクラスを統合するためのグループ正則化項であり，本稿ではグループ lasso (Yuan and Lin, 2006)を用いる．このグループ正則化項により，$\boldsymbol{\beta}_j$ $(j = 1, \ldots, J-1)$ が 0 ベクトルか否かが判定され，もし，$\boldsymbol{\beta}_j = \mathbf{0}$ ならば j 番目のクラスと $(j+1)$ 番目のクラスを統合する．なお，p_j を $\boldsymbol{\beta}_j$ の要素数と定義した場合，グループ lasso は通常 $\sum_{j=1}^{J-1} \sqrt{p_j} \|\boldsymbol{\beta}_j\|_2$ と定式化されるが，(3.1)式においては $p_1 = \cdots = p_{J-1}$ のため，$\sqrt{p_j}$ は含めずに定式化していることに注意されたい．

3.2 推定アルゴリズム

(3.1)式に含まれるグループ正則化項は，微分不可能な点をもつため，ニュートン・ラフソン法等の非線形最適化手法を用いて推定値を求めることは困難である．本稿では，交互方向乗数法(Boyd et al., 2011)に基づいた推定アルゴリズムにより(3.1)式に対するパラメータ推定値を得る．

まず，(3.1)式を最小化問題に書き換えると

$$\min_{\boldsymbol{\beta}} \left\{ -\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{J-1} \|\boldsymbol{\beta}_j\|_2 \right\}$$

となる．ここで，$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_{J-1}^\top)^\top$ である．次に，微分不可能な正則化項を分離するために $(p+1)$ 次元の変数 $\boldsymbol{z}_j$ $(j = 1, \ldots, J-1)$ を用いて，以下の等式制約付きの最小化問題に変形する．

$$\min_{\boldsymbol{\beta}, \boldsymbol{z}} \left\{ -\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{J-1} \|\boldsymbol{z}_j\|_2 \right\} \quad \text{subject to} \quad \boldsymbol{\beta} = \boldsymbol{z}.$$

ここで，$\boldsymbol{z} = (\boldsymbol{z}_1^\top, \ldots, \boldsymbol{z}_{J-1}^\top)^\top$ である．したがって，この最小化問題に対する拡張ラグランジアンは

$$L_\rho(\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{u}) = -\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{J-1} \|\boldsymbol{z}_j\|_2 + \sum_{j=1}^{J-1} \boldsymbol{u}_j^\top (\boldsymbol{\beta}_j - \boldsymbol{z}_j) + \frac{\rho}{2} \sum_{j=1}^{J-1} \|\boldsymbol{\beta}_j - \boldsymbol{z}_j\|_2^2$$

となる．ここで，$\boldsymbol{u}_j$ $(j = 1, \ldots, J-1)$ は $(p+1)$ 次元のラグランジュ乗数で，$\boldsymbol{u} = (\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_{J-1}^\top)^\top$ である．また，ρ は正の値を取る調整パラメータである．

交互方向乗数法では，まず各パラメータ $(\boldsymbol{\beta}, \boldsymbol{z})$ について，拡張ラグランジアンを最小化するパラメータを推定する．次に，その最小となるパラメータを用いてラグランジュ乗数を勾配法(最急上昇法)により更新する．つまり，t $(t = 1, 2, \ldots)$ 回目の更新値をそれぞれ $\boldsymbol{\beta}^t, \boldsymbol{z}_j^t, \boldsymbol{u}^t$ とすると，

$$\boldsymbol{\beta}^{t+1} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ L_\rho(\boldsymbol{\beta}, \boldsymbol{z}^t, \boldsymbol{u}^t), \tag{3.2}$$

$$\boldsymbol{z}_j^{t+1} = \underset{\boldsymbol{z}_j}{\operatorname{argmin}}\ L_\rho(\boldsymbol{\beta}^{t+1}, \boldsymbol{z}_j, \boldsymbol{u}^t), \quad j = 1, \ldots, J-1, \tag{3.3}$$

$$\boldsymbol{u}^{t+1} = \boldsymbol{u}^t + \rho(\boldsymbol{\beta}^{t+1} - \boldsymbol{z}^{t+1})$$

のように更新していく．

(3.2)式と(3.3)式の具体的な更新式を述べる．まず $\boldsymbol{\beta}$ の更新式(3.2)について，右辺の最小化は，

$$\min_{\boldsymbol{\beta}} \left\{ -\ell(\boldsymbol{\beta}) + \sum_{j=1}^{J-1} \boldsymbol{u}_j^{t\top}(\boldsymbol{\beta}_j - \boldsymbol{z}_j^t) + \frac{\rho}{2} \sum_{j=1}^{J-1} \|\boldsymbol{\beta}_j - \boldsymbol{z}_j^t\|_2^2 \right\}$$

と同値である．この最小化問題はパラメータ $\boldsymbol{\beta}$ に対して微分可能であるため，更新値 $\boldsymbol{\beta}^{t+1}$ はニュートン・ラフソン法により得ることができる．ニュートン・ラフソン法を実行するための 1 次導関数および 2 次導関数は順に，

$$\frac{\partial L_\rho}{\partial \boldsymbol{\beta}} = -\frac{\partial \ell}{\partial \boldsymbol{\beta}} + \rho\boldsymbol{\beta} - \rho\boldsymbol{z} + \boldsymbol{u},$$

$$\frac{\partial^2 L_\rho}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} = -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} + \rho I_{(J-1)(p+1)}$$

となる．ここで，$I_{(J-1)(p+1)}$ は $(J-1)(p+1) \times (J-1)(p+1)$ 型の単位行列であり，対数尤度関数の 1 次導関数と 2 次導関数の具体的な式は付録 A に載せている．

次に $\boldsymbol{z}$ の更新式について述べる．(3.3)式について，右辺の最小化は，

$$\min_{\boldsymbol{z}} \left\{ \lambda \sum_{j=1}^{J-1} \|\boldsymbol{z}_j\|_2 + \sum_{j=1}^{J-1} \boldsymbol{u}_j^{t\top}(\boldsymbol{\beta}_j^{t+1} - \boldsymbol{z}_j) + \frac{\rho}{2} \sum_{j=1}^{J-1} \|\boldsymbol{\beta}_j^{t+1} - \boldsymbol{z}_j\|_2^2 \right\} \tag{3.4}$$

と同値である．ここで，等式

$$\boldsymbol{u}_j^{t\top}(\boldsymbol{\beta}_j^{t+1} - \boldsymbol{z}_j) + \frac{\rho}{2} \left\|\boldsymbol{\beta}_j^{t+1} - \boldsymbol{z}_j\right\|_2^2 = \frac{\rho}{2} \left\|\boldsymbol{\beta}_j^{t+1} - \boldsymbol{z}_j + \frac{1}{\rho}\boldsymbol{u}_j^t\right\|_2^2 - \frac{\rho}{2} \left\|\frac{1}{\rho}\boldsymbol{u}_j^t\right\|_2^2$$

に注意すると，問題(3.4)の $\boldsymbol{z}$ に関する最小化は，

$$\sum_{j=1}^{J-1} \rho \left[\frac{1}{2} \left\| \boldsymbol{z}_j - \left\{ \boldsymbol{\beta}_j^{t+1} + \frac{1}{\rho}\boldsymbol{u}_j^t \right\} \right\|_2^2 + \frac{\lambda}{\rho} \|\boldsymbol{z}_j\|_2 \right]$$

の $\boldsymbol{z}$ に関する最小化と同値となる．またこれは，各 j $(j = 1, \ldots, J-1)$ 毎に問題が分離されているため，更新式は

$$\boldsymbol{z}_j^{t+1} = S_{\frac{\lambda}{\rho}} \left(\boldsymbol{\beta}_j^{t+1} + \frac{1}{\rho}\boldsymbol{u}_j^t \right), \quad j = 1, \ldots, J-1 \tag{3.5}$$

となる．ここで，$S_v(\boldsymbol{a})$ はベクトルに対する軟閾値作用素，

$$S_v(\boldsymbol{a}) = \left(1 - \frac{v}{\|\boldsymbol{a}\|_2}\right)_+ \boldsymbol{a}$$

であり，$(\cdot)_+$ は任意の実数 x に対して，$(x)_+ = \max(0, x)$ と定義される．なお，更新式(3.5)の導出については川野 他 (2018)を参照されたい．

3.3 正則化パラメータの選択

パラメータの推定値は，調整パラメータ ρ と正則化パラメータ λ を含んでいる．調整パラメータ ρ は Boyd et al. (2011) にしたがい $\rho = 1$ と設定する．一方，正則化パラメータ λ は情報量規準 BIC (Schwarz, 1978) に基づいて選択する．

BIC はその第 2 項目にモデルの自由度を必要とする．Yuan and Lin (2006) は，線形回帰モデルにおいて，グループ lasso により推定したときの推定値 $\hat{\boldsymbol{\xi}}$ を持つモデルの自由度を

$$\sum_{j=1}^{J} I(\|\hat{\boldsymbol{\xi}}_j\|_2 > 0) + \sum_{j=1}^{J} \frac{\|\hat{\boldsymbol{\xi}}_j\|_2}{\|\hat{\boldsymbol{\xi}}_j^{\mathrm{LS}}\|_2}(p_j - 1)$$

と定義した．ここで $\hat{\boldsymbol{\xi}}_j^{\mathrm{LS}}$ は最小二乗推定値である．本稿ではこの考えに基づいて，提案手法のモデルの自由度を

$$\mathrm{df} = \sum_{j=1}^{J-1} I(\|\hat{\boldsymbol{\beta}}_j\|_2 > 0) + p \sum_{j=1}^{J-1} \frac{\|\hat{\boldsymbol{\beta}}_j\|_2}{\|\hat{\boldsymbol{\beta}}_j^{\mathrm{ridge}}\|_2} \tag{3.6}$$

とする．ここで，$\hat{\boldsymbol{\beta}}_j$ は最大化問題 (3.1) の解，$\hat{\boldsymbol{\beta}}_j^{\mathrm{ridge}}$ はリッジ推定値を表している．リッジ推定値に含まれる正則化パラメータの値は 10^{-5} に固定した．最尤推定値ではなくリッジ推定値を用いた理由は，4 節の数値実験において最尤推定値が求まらなかったことが多々あったためである．また，$p_1 = \cdots = p_{J-1} = p + 1$ となる関係性に注意されたい．

以上より，本稿で用いる情報量規準 BIC は (3.6) 式の自由度を用いて

$$\mathrm{BIC} = -2\ell(\hat{\boldsymbol{\beta}}) + \mathrm{df}\log(n)$$

となる．BIC の値が最小となる正則化パラメータ λ の値を最適な値として採用する．

3.4 関連研究

Price et al. (2019) は，順序関係のないカテゴリカルデータを目的変数とした多項ロジットモデルに対し，クラス統合を実行する方法として GFMR を提案した．GFMR は以下の最小化問題として定式化される．

$$\min_{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_J} \left\{ -\ell_{\mathrm{MLM}}(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{J-1}) + \lambda \sum_{(j,m)\in\mathcal{L}} \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_m\|_2 \right\} \quad \text{subject to} \quad \boldsymbol{\beta}_J = \mathbf{0}.$$

ここで，$\ell_{\mathrm{MLM}}(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{J-1})$ は多項ロジットモデルに対する対数尤度関数，λ は正の値をとる正則化パラメータである．第 2 項目はクラス統合を行うための正則化項であり，候補集合 $\mathcal{L}$ は $\mathcal{S} = \{(a,b) \in \mathcal{J}\times\mathcal{J} : a < b\}$ からクラスの組み合わせを任意に設定する．ここで $\mathcal{J} = \{1,\ldots,J\}$ である．例えば，1 番目のクラスと 3 番目のクラスが統合する可能性がある場合は，$\mathcal{L} = \{(1,3)\}$ とする．クラスの関係性に何も事前の情報がなければ，$\mathcal{L} = \mathcal{S}$ とする．$\boldsymbol{\beta}_J = \mathbf{0}$ は多項ロジットモデルに関する制約条件である．なお，クラスを統合する方法として，提案手法ではグループ lasso，GFMR ではグループ連結 lasso (group fused lasso) (Bleakley and Vert, 2011; Alaíz et al., 2013) を用いていることに注意されたい．

推定アルゴリズムは，交互方向乗数法により得ている．更新式の詳細については Price et al. (2019) を参照されたい．正則化パラメータの値は，K 分割交差検証法により選択している．具体的には，K 分割したデータセットのうち k 番目のデータセットを $\mathcal{V}_k$ としたとき，評価関数

$$\sum_{k=1}^{K} \sum_{i\in\mathcal{V}_k} \sum_{l=1}^{J} y_{il} \log(\hat{\pi}_l^{(-\mathcal{V}_k)}(\boldsymbol{x}_i, \lambda)) \tag{3.7}$$

を最大にする正則化パラメータの値を採用している．ここで，y_{il} は(2.2)と同様にデータ i がクラス l に属しているときに 1，その他のときに 0 をとる指示変数である．$\hat{\pi}_l^{(-\nu_k)}(\boldsymbol{x}_i, \lambda)$ は，k 番目のデータセットを除いたデータセットにより推定したモデルの事後確率を表す．

4. 数値実験

本節では，モンテカルロ・シミュレーションと実データへの適用を通して，提案手法の有効性を検証する．

4.1 モンテカルロ・シミュレーション

真のモデルを

$$\log \frac{\pi_j^*(\boldsymbol{x}_i^\dagger)}{\pi_{j+1}^*(\boldsymbol{x}_i^\dagger)} = \boldsymbol{\beta}_j^{*\top} \boldsymbol{x}_i, \quad (i = 1, \ldots, n,\ j = 1, \ldots, J-1) \tag{4.1}$$

と設定した．ここで，回帰係数パラメータ $\boldsymbol{\beta}_j^* = (\beta_{j0}^*, \beta_{j1}^*, \ldots, \beta_{j10}^*)^\top$ の設定は次の 8 通りとした．

- 設定 1：$(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*) = (\mathbf{0}, \boldsymbol{\delta}, \mathbf{0})$
- 設定 2：$(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*) = (\mathbf{0}, \boldsymbol{\delta}, \boldsymbol{\delta})$
- 設定 3：$(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*, \boldsymbol{\beta}_4^*) = (\mathbf{0}, \boldsymbol{\delta}, \boldsymbol{\delta}, \boldsymbol{\delta})$
- 設定 4：$(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*, \boldsymbol{\beta}_4^*) = (\mathbf{0}, \boldsymbol{\delta}, \boldsymbol{\delta}, \mathbf{0})$
- 設定 5：$(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*)^\top = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0.5 & 1 & 0.5 & 1 & 0.5 & 1 & 0.5 & 1 & 0.5 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
- 設定 6：$(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*)^\top = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0.5 & 0.5 & 0.5 \\ 1 & 1 & 1 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \end{pmatrix}$
- 設定 7：$(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*, \boldsymbol{\beta}_4^*)^\top = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0.5 & 0.5 & 0.5 \\ 1 & 1 & 1 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 1 & 0.5 & 1 & 0.5 & 1 & 0.5 & 1 & 0.5 & 1 & 0.5 & 1 \end{pmatrix}$
- 設定 8：$(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*, \boldsymbol{\beta}_4^*)^\top = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0.5 & 0.5 & 0.5 \\ 1 & 1 & 1 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

ただし，$\boldsymbol{\delta} = (\underbrace{\delta, \ldots, \delta}_{11})$ であり，$\delta = 0.5, 1$ の 2 通りを考えた．

ここで，設定したパラメータの意味について，設定 1 を例に取って説明する．設定 1 では $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_3^* = \mathbf{0}$ であることから，クラス 1 とクラス 2，クラス 3 とクラス 4 はそれぞれ事後確率が等しいことがわかる．すると，3.1 節の議論より，クラス 1 とクラス 2，クラス 3 とクラス 4 はそれぞれ同じクラスであり，実際には 2 クラスしかないことがわかる．つまり，設定 1 は，見かけ上は 4 クラスであるが，実際は 2 クラスしかないシミュレーション設定である．これより，すべての設定における実際のクラスの個数と見かけ上のクラスの個数は，表 1 のようにまとめることができる．

表 1. 実際のクラスの個数と見かけ上のクラスの個数.

	設定 1	設定 2	設定 3	設定 4	設定 5	設定 6	設定 7	設定 8
実際のクラスの個数	2	3	4	3	2	3	4	3
見かけ上のクラスの個数	4	4	5	5	4	4	5	5

説明変数に関するデータ $\boldsymbol{x}_i^\dagger$ は $N(\mathbf{0}, \Sigma)$ から発生させた．ここで，Σ は 10×10 型の分散共分散行列であり，その (i, j) 成分を $\zeta^{|i-j|}$ とし，$\zeta = 0, 0.5$ の 2 通りを考えた．(4.1)式より計算される真の事後確率 $\pi_j^*(\boldsymbol{x}_i^\dagger)$ $(j = 1, \ldots, J-1)$ から，乱数を用いて各データ $\boldsymbol{x}_i^\dagger$ $(i = 1, \ldots, n)$ が所属するクラスを決定した．サンプルサイズは $n = 200, 300, 500$ とし，シミュレーション回数は 100 回とした．

比較手法として，Price et al. (2019)による GFMR と 2 節で述べた ACL を用いた．ACL に含まれるパラメータはリッジ推定により求め，正則化パラメータの値は 10^{-5} に固定した．提案手法と GFMR に含まれる正則化パラメータの値は，3.3 節の BIC により選択した．正則化パラメータの値の候補はベイズ最適化(Snoek et al., 2012)を用い，提案手法では 0.1 から 7，GFMR では 0.1 から 100 の範囲で探索した．ベイズ最適化の計算には統計ソフトウェア R のパッケージ **rBayesianOptimization** を用いた．GFMR の候補集合 $\mathcal{L}$ については，順序付きカテゴリカルデータを解析する際に Price et al. (2019)で推奨されている $\mathcal{L} = \{(1, 2), (2, 3), \ldots, (J-1, J)\}$ を用いた．

クラス統合後のクラスの個数，統合正答回数，真の事後確率と推定した事後確率間のカルバック・ライブラー情報量(KL)の 3 つの観点からそれぞれの手法を評価した．100 回のシミュレーションにおいて，クラスが正しく統合された回数を，統合正答回数とした．ここで，各シミュレーションにおいてクラスが正しく統合されるとは，見かけ上のクラスが実際のクラスに統合されることをいう．例えば，設定 1 においては，クラス 1 とクラス 2，クラス 3 とクラス 4 がそれぞれ統合されるときは正しく統合されているといえるが，クラス 1 とクラス 2，クラス 2 とクラス 3 がそれぞれ統合されるときは正しく統合されているとはいえない．また，KL は

$$\mathrm{KL}(\hat{\pi}, \pi^*) = \sum_{j=1}^{J} \log \left(\frac{\hat{\pi}_j(\boldsymbol{x}_i^\dagger, \lambda)}{\pi_j^*(\boldsymbol{x}_i^\dagger)} \right) \hat{\pi}_j(\boldsymbol{x}_i^\dagger, \lambda)$$

と定義される．ただし，$\hat{\pi}_j(\boldsymbol{x}_i^\dagger, \lambda)$ は推定したモデルから求めた事後確率である．100 回のシミュレーション毎にテストデータを 100 個発生させて KL を計算し，得られた KL の平均と標準偏差を計算した．

表 2，表 3，表 4 はそれぞれ $\zeta = 0$ とした場合の設定 1 と設定 2，設定 3 と設定 4，設定 5 から設定 8 に対する計算結果である．また，表 5，表 6，表 7 はそれぞれ $\zeta = 0.5$ とした場合の設定 1 と設定 2，設定 3 と設定 4，設定 5 から設定 8 に対する計算結果である．なお，設定 3 の $n = 200, \delta = 1, \zeta = 0.5$ においては，リッジ推定値が発散して計算することができなかった．

まず，$\zeta = 0$ のときの結果を考察する．表 2，表 3，表 4 から，ほぼすべての場合において，提案手法は ACL よりも小さい KL の値を与えており，よりあてはまりの良いモデルを構築できていることがわかる．また，GFMR と比較しても，同様に，提案手法がよりあてはまりの良いモデルを構築できていることがわかる．統合正答回数の観点から見ると，サンプルサイズが 200 の場合には GFMR が他の手法よりも多いが，サンプルサイズが大きくなるにつれ提案手法が多くなる．

次に，$\zeta = 0.5$ のときの結果を考察する．表 5，表 6，表 7 から，$\zeta = 0$ のときに比べ GFMR

表 2. 設定 1 から設定 2 に対する結果($\zeta = 0$)．実験の各設定において，KL が一番小さい値，統合正答回数が一番多い値を太字で表している．

				提案手法			GFMR			ACL
	n	δ		クラス個数	KL	統合正答回数	クラス個数	KL	統合正答回数	KL
設定 1	200	0.5	mean	1.58	34.50	31	1.97	**10.86**	**97**	17.72
			sd	0.74	23.27		0.17	8.80		4.66
		1	mean	2.69	**11.51**	51	2.08	11.76	**92**	18.84
			sd	0.83	15.76		0.27	5.07		4.90
	300	0.5	mean	1.96	21.53	71	1.93	**21.37**	**76**	17.79
			sd	0.57	27.91		0.57	27.98		3.68
		1	mean	2.16	**9.35**	87	2.09	12.55	**92**	19.00
			sd	0.44	4.03		0.32	5.59		4.34
	500	0.5	mean	2.02	**8.77**	**99**	2.15	24.67	73	17.39
			sd	0.20	3.74		0.78	42.08		4.47
		1	mean	2.00	**8.94**	**100**	2.04	11.33	96	17.67
			sd	0.00	4.13		0.20	4.95		4.66
設定 2	200	0.5	mean	3.30	**16.17**	58	2.66	18.95	**66**	18.78
			sd	0.67	21.99		0.48	8.24		4.32
		1	mean	3.11	**16.14**	89	3.02	22.69	**96**	18.90
			sd	0.31	4.14		0.20	10.00		4.69
	300	0.5	mean	3.09	**11.42**	**91**	2.87	16.91	87	18.10
			sd	0.29	3.94		0.34	8.60		4.39
		1	mean	3.01	**15.60**	**99**	3.02	23.65	98	19.08
			sd	0.10	4.40		0.14	11.66		4.69
	500	0.5	mean	3.01	**12.28**	**99**	2.81	40.43	80	17.79
			sd	0.10	3.97		0.66	84.28		4.19
		1	mean	3.00	**14.66**	100	3.00	25.67	100	17.32
			sd	0.00	3.51		0.00	15.40		3.80

はより良い結果を与えているが，総合的に見て提案手法の方が KL の値は小さい．また，統合正統回数もサンプルサイズが大きくなるにつれ提案手法が GFMR より大きい値をとっていることがわかる．以上より，提案手法は従来手法よりも KL を小さくし，さらにサンプルサイズが大きい場合は正確にクラス統合を実行することがわかる．

提案手法において BIC により選択されるモデルと K 分割交差検証法(CV)により選択されるモデルとを比較した．交差検証法には(3.7)式を用い，分割数は $K = 5$ とした．KL と統合正答回数の 2 つの観点から評価し，KL については BIC による結果を分子，CV による結果を分母とし，統合正答回数については CV による結果を分子，BIC による結果を分母としてそれらの割合を計算した．これらより，KL と統合正答回数のどちらの指標においても，1 より小さい値の場合は本稿で用いた BIC の方がより良いといえる．

設定 1 から設定 4 の結果を表 8，設定 5 から設定 8 の結果を表 9 に載せている．KL については，ときおり 1 より大きい値をとっているが，基本的には 1 より小さい．統合正答回数については，ほとんどすべての場合において，1 より小さい値をとっている．したがって，BIC の方が KL を小さくし，統合正答回数を多くするモデルを選択することがわかる．

4.2 実データへの適用

赤ワインの品質ランクデータ(Cortez et al., 2009)に対し提案手法を適用した．本データは，

表 3. 設定 3 から設定 4 に対する結果($\zeta = 0$)．実験の各設定において，KL が一番小さい値，統合正答回数が一番多い値を太字で表している．

				提案手法			GFMR			ACL
	n	δ		クラス個数	KL	統合正答回数	クラス個数	KL	統合正答回数	KL
設定 3	200	0.5	mean	4.45	**16.29**	**55**	3.22	31.27	35	25.39
			sd	0.50	4.19		0.75	13.93		5.89
		1	mean	4.22	**29.17**	78	3.91	43.80	**85**	29.27
			sd	0.42	5.36		0.38	17.99		7.23
	300	0.5	mean	4.15	**16.07**	**85**	3.48	30.73	55	24.01
			sd	0.36	3.90		0.63	16.36		4.81
		1	mean	4.01	**24.32**	**99**	3.78	37.81	74	26.86
			sd	0.10	4.77		0.46	17.70		5.91
	500	0.5	mean	4.01	**16.63**	**99**	3.70	26.62	68	23.93
			sd	0.10	4.72		0.48	14.78		5.45
		1	mean	4.00	**23.45**	**100**	3.86	38.93	86	24.50
			sd	0.00	4.69		0.35	18.53		4.58
設定 4	200	0.5	mean	3.53	41.90	18	2.39	**21.63**	**39**	24.77
			sd	1.49	53.07		0.49	8.01		5.82
		1	mean	3.87	**20.73**	38	3.00	27.05	**86**	28.35
			sd	0.79	4.22		0.38	12.14		6.19
	300	0.5	mean	3.64	**18.69**	51	2.69	24.22	**71**	24.67
			sd	0.88	28.73		0.51	28.92		5.15
		1	mean	3.12	**17.63**	90	3.00	25.31	**94**	26.86
			sd	0.38	5.52		0.25	13.42		6.17
	500	0.5	mean	3.13	**14.17**	**87**	3.03	41.57	61	23.68
			sd	0.39	7.04		0.97	82.46		5.98
		1	mean	3.01	**16.04**	99	3.01	27.82	99	23.47
			sd	0.10	4.76		0.10	11.44		5.27

UCI Machine Learning Repository(`https://archive.ics.uci.edu/ml/index.php`)から取得できる．赤ワインの品質は 6 段階で評価されており，段階が上がるにつれて赤ワインの品質の評価は高くなる．つまり，1 段階目の赤ワインの品質が最も評価が低く，6 段階目に属する赤ワインの品質が最も評価が高い．この評価された段階を目的変数とし，以降段階をクラスと呼ぶ．説明変数には，酒石酸濃度や残留糖分濃度などの赤ワインの成分に対する 11 種類の変数を用いた．各クラスのサンプルサイズは表 10 の通りである．

比較手法として GFMR を考え，BIC により正則化パラメータの値を選択した．ここでは(3.6)式の自由度に含まれるリッジ推定値が発散したため，自由度 $\mathrm{df} = (p+1)\left(\sum_{j=1}^{J-1} I(\|\hat{\beta}_j\|_2 > 0)\right)$ をもつ BIC を用いた．正則化パラメータの値の候補は，提案手法では 0.1 から 7，GFMR では 0.1 から 200 の範囲からベイズ最適化を用いて探索した．GFMR の候補集合 $\mathcal{L}$ については，4.1 節と同様に $\mathcal{L} = \{(1,2),(2,3),\ldots,(J-1,J)\}$ を用いた．

ブートストラップ法(Efron and Tibshirani, 1993)を用いてクラス統合の安定性を検証した．まずクラス 1 をクラス 1a とクラス 1b に，サンプルサイズが同じになるようランダムに 2 分割し，このデータセットからブートストラップ標本を発生させて提案手法と GFMR を当てはめた．この操作を 100 回繰り返し，クラス 1a とクラス 1b が統合された回数(統合正答回数)を数え上げた．なお，統合正答回数の数え上げでは，各ブートストラップ標本においてクラス 1a とクラス 1b だけを統合するときに限って数え上げることに注意しておく．クラス 2 からクラス

表 4. 設定 5 から設定 8 の結果($\zeta = 0$)．実験の各設定において，KL が一番小さい値，統合正答回数が一番多い値を太字で表している．

			提案手法			GFMR			ACL
	n		クラス個数	KL	統合正答回数	クラス個数	KL	統合正答回数	KL
設定 5	200	mean	2.72	**10.95**	52	2.03	12.05	**97**	18.82
		sd	0.87	13.29		0.17	4.743		5.12
	300	mean	2.11	**8.77**	91	2.05	11.19	**95**	18.20
		sd	0.37	3.76		0.22	4.87		4.56
	500	mean	2.00	**8.72**	**100**	2.04	11.52	97	17.55
		sd	0.00	3.99		0.24	4.27		4.71
設定 6	200	mean	3.14	**13.66**	**86**	2.72	22.59	70	20.10
		sd	0.35	3.84		0.47	9.76		5.53
	300	mean	3.00	**12.65**	**100**	2.78	22.76	78	18.53
		sd	0.00	3.62		0.42	10.75		4.57
	500	mean	3.00	**12.83**	**100**	2.96	18.38	96	17.43
		sd	0.00	4.22		0.20	10.11		4.75
設定 7	200	mean	4.24	**20.49**	**76**	3.65	35.30	67	26.07
		sd	0.43	4.69		0.63	15.60		6.86
	300	mean	4.00	**18.44**	**100**	3.79	32.95	79	24.52
		sd	0.00	3.95		0.41	15.53		5.54
	500	mean	4.00	**19.09**	**100**	3.83	32.26	83	23.85
		sd	0.00	4.10		0.38	13.32		5.89
設定 8	200	mean	4.09	**15.73**	29	2.60	24.82	**56**	26.02
		sd	0.82	4.30		0.53	9.57		5.69
	300	mean	3.22	**14.54**	**80**	2.78	21.80	76	24.85
		sd	0.46	4.36		0.44	9.14		5.35
	500	mean	3.05	**14.65**	**95**	2.93	18.63	93	23.63
		sd	0.22	4.67		0.26	10.24		5.94

6 に対しても同様の手順でブートストラップ標本を発生させ，統合正答回数を数え上げた．

結果を表 11 にまとめている．クラス 2 とクラス 5 以外の結果については，提案手法は統合正答回数が 80 回以上であり，GFMR より多くなっている．特に，クラス 1 とクラス 6 の結果においては，提案手法の統合正答回数は GFMR の 2 倍以上である．一方，クラス 2 やクラス 5 では，そもそも統合正答回数が少ない．したがって，このときの結果についてはあまり議論すべきではないかも知れないが，クラス 2 においては提案手法の統合正答回数は GFMR より多くなっている．以上より，統合正答回数の観点から，提案手法は GFMR よりも多くの場合において安定してクラス統合が行われていることがわかる．

5. まとめ

隣接カテゴリーロジットモデルとグループ lasso により，順序付きカテゴリカルデータを対象とするモデルにおけるクラス統合の方法を提案した．推定アルゴリズムを交互方向乗数法により構成し，グループ lasso により推定されたモデルの自由度をもつ情報量規準 BIC により正則化パラメータの値を選択した．数値実験の結果をいくつかの観点から考察することにより，従来手法に比べて提案手法は多くの点で有用であることがわかった．

4.1 節のモンテカルロ・シミュレーションにおいて，サンプルサイズが小さいときに，提案

表 5. 設定 1 から設定 2 の結果($\zeta = 0.5$). 実験の各設定において，KL が一番小さい値，統合正答回数が一番多い値を太字で表している．

				提案手法			GFMR			ACL
	n	δ		クラス個数	KL	統合正答回数	クラス個数	KL	統合正答回数	KL
設定 1	200	0.5	mean	2.95	16.83	33	2.01	**8.00**	**99**	18.91
			sd	1.02	26.89		0.10	4.21		5.13
		1	mean	2.47	**8.42**	68	2.08	11.45	**93**	20.18
			sd	0.74	3.55		0.31	5.35		5.20
	300	0.5	mean	2.67	**9.45**	59	2.12	10.36	**88**	18.43
			sd	0.87	5.05		0.41	16.67		4.62
		1	mean	2.06	**7.19**	**95**	2.08	10.32	92	18.57
			sd	0.28	3.62		0.27	4.78		4.83
	500	0.5	mean	2.03	**7.24**	**98**	2.36	28.47	69	17.07
			sd	0.22	3.50		0.87	70.34		4.28
		1	mean	2.00	**7.65**	**100**	2.05	10.11	95	17.84
			sd	0.00	3.62		0.22	4.56		5.12
設定 2	200	0.5	mean	3.51	**10.76**	49	2.99	14.95	**95**	19.69
			sd	0.50	3.46		0.22	9.81		5.30
		1	mean	3.18	**16.73**	82	3.02	23.00	**98**	21.77
			sd	0.39	3.93		0.14	10.56		6.26
	300	0.5	mean	3.21	**10.94**	79	2.98	14.50	**96**	18.94
			sd	0.41	4.10		0.20	10.91		4.98
		1	mean	3.03	**13.72**	97	3.03	20.50	97	19.99
			sd	0.17	3.47		0.17	10.40		5.68
	500	0.5	mean	3.03	**11.00**	**97**	3.02	28.11	92	18.35
			sd	0.17	4.09		0.38	84.41		5.01
		1	mean	3.00	**13.60**	100	3.00	23.85	100	18.19
			sd	0.00	3.49		0.00	12.82		4.67

手法が GFMR に統合正答回数で劣ることがあった．この理由としては，次の 2 点が考えられる．1 点目は，サンプルサイズとクラス間の順序との関係性である．一般に，サンプルサイズが大きい場合には各クラスの性質が明確になるため，クラス間の順序が強調されることが知られている．一方，サンプルサイズが小さい場合には各クラスの性質が明確にならないため，クラス間の順序が強調されず，提案手法の統合正答回数は少なくなったと考えられる．2 点目は，正則化パラメータの選択方法である．サンプルサイズが小さい場合には，統合後のクラスの個数が真のクラスの個数よりも大きくなり，その結果統合正答回数が少なくなる状況が見受けられた．これは，正則化パラメータの値が小さめに選択されていることを表しており，BIC を用いた選択方法がうまく機能していないことを示唆している．

4.2 節の実データへの適用において，クラス 2 とクラス 5 の統合正答回数が非常に少なかった．これは次の 2 つの理由が重なったため生じたと考えられる．1 つ目は，クラス 2 とクラス 5 はそれぞれ 2 つのクラスに挟まれており，統合する隣接クラスの候補が多いということ．2 つ目は，クラス 3 とクラス 4 に比べて，サンプルサイズが小さいということである．このような状況において，統合正答回数を改善する統計モデルの構築に取り組むことは大変意義がある．

本稿で用いた BIC におけるモデルの自由度は Yuan and Lin (2006) の考えに基づいて導出したが，Yuan and Lin (2006) で導出されている自由度は線形回帰モデルかつ計画行列の各列が直交するという仮定が課されている．この点において提案モデルと乖離しており，提案モデルの

表 6. 設定 3 から設定 4 の結果($\zeta = 0.5$). 実験の各設定において，KL が一番小さい値，統合正答回数が一番多い値を太字で表している.

				提案手法			GFMR			ACL
	n	δ		クラス個数	KL	統合正答回数	クラス個数	KL	統合正答回数	KL
設定 3	200	0.5	mean	4.54	**15.70**	46	3.65	29.13	**66**	27.11
			sd	0.50	4.17		0.50	17.49		7.33
		1	mean	–	–	–	–	–	–	–
			sd	–	–		–	–		–
	300	0.5	mean	4.29	**14.51**	**71**	3.49	23.84	49	26.21
			sd	0.46	4.28		0.50	12.94		6.48
		1	mean	4.09	**26.10**	**91**	3.72	48.24	64	28.23
			sd	0.29	5.37		0.53	26.82		5.44
	500	0.5	mean	4.06	**15.10**	**94**	3.65	27.45	65	24.32
			sd	0.24	4.33		0.48	15.35		5.46
		1	mean	4.00	**21.39**	**100**	3.50	44.59	48	25.22
			sd	0.00	4.06		0.52	18.69		6.07
設定 4	200	0.5	mean	4.57	**14.02**	11	2.77	20.74	**75**	27.09
			sd	0.69	4.03		0.45	14.32		6.75
		1	mean	3.75	**20.85**	53	3.07	27.91	**93**	29.66
			sd	0.87	4.90		0.26	9.68		6.95
	300	0.5	mean	3.67	**12.56**	52	3.00	14.96	**94**	25.32
			sd	0.78	4.49		0.25	9.13		5.58
		1	mean	3.19	**14.83**	84	3.03	23.54	**97**	26.43
			sd	0.46	4.68		0.17	12.98		5.79
	500	0.5	mean	3.14	**12.16**	91	3.12	20.96	**92**	23.56
			sd	0.47	4.34		0.56	67.52		4.77
		1	mean	3.00	**14.37**	**100**	3.04	28.18	96	24.47
			sd	0.00	4.04		0.20	13.90		5.64

自由度についてはまだ多くの問題が残っている．以上は今後の研究課題としたい.

謝　辞

本稿の改訂に当たって，編集委員および査読者の方から大変貴重なご指摘と適切なご意見をいただきました．ここに記して御礼申し上げます．本研究は日本学術振興会科学研究費補助金(19K11854)の助成を受けたものです.

付　録

A. 隣接カテゴリーロジットモデルの対数尤度関数の 1 次導関数および 2 次導関数

(2.3)式の隣接カテゴリーロジットモデルの対数尤度関数は,

$$\ell(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{J-1}) = \sum_{i=1}^{n} \left\{ \mathbf{1}_{J-1}^{\top} Y_i^{**} X_i^{*} \boldsymbol{\beta} - \log \left[1 + \sum_{j=1}^{J-1} \exp\left(\boldsymbol{I}_j^{\top} X_i^{*} \boldsymbol{\beta} \right) \right] \right\} \tag{A.1}$$

と書き換えることができる．ここで，$\mathbf{1}_{J-1}$ は要素がすべて 1 の $(J-1)$ 次元ベクトル，$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\top}, \ldots, \boldsymbol{\beta}_{J-1}^{\top})^{\top}$ は $(J-1)(p+1)$ 次元ベクトル，$\boldsymbol{I}_j = (I(1 > j-1), \ldots, I(J-1 > j-1))^{\top}$

表 7. 設定 5 から設定 8 の結果($\zeta = 0.5$)．実験の各設定において，KL が一番小さい値，統合正答回数が一番多い値を太字で表している．

			提案手法			GFMR			ACL
	n		クラス個数	KL	統合正答回数	クラス個数	KL	統合正答回数	KL
設定 5	200	mean	2.71	**8.82**	54	2.11	11.41	**90**	19.66
		sd	0.84	3.98		0.35	4.76		5.70
	300	mean	2.12	**7.76**	90	2.03	10.76	**97**	18.74
		sd	0.38	4.01		0.17	5.20		5.34
	500	mean	2.00	**7.10**	**100**	2.06	10.24	94	17.51
		sd	0.00	3.00		0.24	3.81		4.41
設定 6	200	mean	3.27	**11.78**	73	3.00	18.48	**96**	19.91
		sd	0.45	3.33		0.20	8.68		4.68
	300	mean	3.08	**12.03**	92	3.01	18.19	**99**	19.38
		sd	0.27	4.29		0.10	6.26		5.56
	500	mean	3.00	**11.81**	100	3.00	17.84	100	18.20
		sd	0.00	3.58		0.00	7.04		4.35
設定 7	200	mean	4.47	**21.99**	53	3.80	33.77	**72**	30.34
		sd	0.50	4.84		0.49	13.62		7.04
	300	mean	4.08	**17.13**	**92**	3.55	31.79	55	26.21
		sd	0.27	4.32		0.50	12.12		5.70
	500	mean	4.01	**17.41**	**99**	3.63	34.64	61	25.76
		sd	0.10	3.97		0.51	13.59		5.71
設定 8	200	mean	3.87	**13.88**	43	2.90	22.75	**73**	26.54
		sd	0.85	4.13		0.54	12.83		5.53
	300	mean	3.30	**13.00**	77	2.98	20.17	**92**	25.58
		sd	0.59	4.59		0.28	11.79		6.37
	500	mean	3.01	**12.36**	99	2.99	17.32	99	23.87
		sd	0.10	3.75		0.10	9.91		5.18

は $(J-1)$ 次元ベクトル，Y_i^{**} は

$$Y_i^{**} = \begin{pmatrix} y_{i1} & y_{i1} & \cdots & y_{i1} \\ 0 & y_{i2} & \cdots & y_{i2} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & y_{iJ-1} \end{pmatrix}$$

となる $(J-1)\times(J-1)$ 型の正方行列，X_i^* は

$$X_i^* = \begin{pmatrix} \boldsymbol{x}_i^\top & 0\cdots 0 & \cdots & 0 \\ 0\cdots 0 & \boldsymbol{x}_i^\top & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0\cdots 0 & 0\cdots 0 & \cdots & \boldsymbol{x}_i^\top \end{pmatrix}$$

となる $(J-1)\times(J-1)(p+1)$ 型の行列である．このとき，(A.1)式の対数尤度関数の 1 次導関数および 2 次導関数は，順に

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left\{ \left(\mathbf{1}_n^\top Y_i^{**} X_i^* \right)^\top - \frac{1}{1+\sum_{j=1}^{J-1}\eta_{ij}} \sum_{j=1}^{J-1} \eta_{ij} \boldsymbol{\gamma}_{ij}^\top \right\},$$

表 8．設定 1 から設定 4 における正則化パラメータ選択法の比較．実験の各設定において，1 より小さい値を太字で表している．

			設定 1		設定 2		設定 3		設定 4	
ζ	n	δ	KL	統合正答回数	KL	統合正答回数	KL	統合正答回数	KL	統合正答回数
0	200	0.5	2.629	**0.677**	1.183	**0.828**	**0.802**	**0.855**	2.526	**0.778**
		1	1.104	**0.824**	**0.991**	**0.685**	1.048	**0.603**	1.006	**0.579**
	300	0.5	1.990	**0.817**	**0.730**	**0.846**	**0.711**	**0.918**	1.199	**0.765**
		1	**0.946**	**0.851**	1.062	**0.828**	1.098	**0.818**	1.006	**0.800**
	500	0.5	**0.854**	**0.788**	**0.782**	**0.919**	**0.965**	**0.909**	**0.916**	**0.874**
		1	**0.999**	**0.950**	1.148	**0.980**	1.168	**0.910**	1.103	**0.899**
0.5	200	0.5	1.579	**0.545**	**0.811**	**0.878**	1.004	**0.739**	**0.724**	1.000
		1	**0.894**	**0.691**	**0.996**	**0.720**	–	–	–	–
	300	0.5	**0.928**	**0.831**	**0.735**	**0.911**	**0.537**	**0.859**	**0.720**	**0.788**
		1	**0.946**	**0.895**	1.026	**0.804**	1.297	**0.802**	**0.992**	**0.798**
	500	0.5	**0.923**	**0.939**	**0.953**	**0.959**	**0.330**	**0.894**	**0.951**	**0.945**
		1	**0.995**	**0.960**	1.105	**0.940**	1.134	**0.940**	1.081	**0.940**

表 9．設定 5 から設定 8 における正則化パラメータ選択法の比較．実験の各設定において，1 より小さい値を太字で表している．

		設定 5		設定 6		設定 7		設定 8	
ζ	n	KL	統合正答回数	KL	統合正答回数	KL	統合正答回数	KL	統合正答回数
0	200	1.044	**0.692**	**0.989**	**0.814**	**0.980**	**0.803**	**0.993**	1.138
	300	**0.975**	**0.901**	**0.990**	**0.900**	1.020	**0.880**	**0.989**	**0.800**
	500	**0.955**	**0.930**	1.006	**0.970**	1.012	**0.910**	**0.986**	**0.895**
0.5	200	**0.940**	**0.815**	**0.954**	**0.712**	1.064	**0.811**	**0.743**	**0.698**
	300	**0.920**	**0.878**	**0.977**	**0.870**	1.032	**0.859**	**0.938**	**0.727**
	500	**0.948**	**0.950**	**0.994**	**0.960**	1.011	**0.919**	**0.993**	**0.949**

表 10．各クラスのサンプルサイズ．

	クラス 1	クラス 2	クラス 3	クラス 4	クラス 5	クラス 6
サンプルサイズ	10	53	681	638	199	18

表 11．各クラスのブートストラップ選択回数．各クラスにおいて統合正答回数が多い方を太字にしている．

	クラス 1	クラス 2	クラス 3	クラス 4	クラス 5	クラス 6
提案手法	**80**	**5**	**100**	**98**	14	**81**
GFMR	38	2	87	79	**20**	37

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = \sum_{i=1}^{n} \left[\frac{1}{\left(1 + \sum_{j=1}^{J-1} \eta_{ij}\right)^2} \left(\sum_{j=1}^{J-1} \eta_{ij} \boldsymbol{\gamma}_{ij} \right) \left(\sum_{j=1}^{J-1} \eta_{ij} \boldsymbol{\gamma}_{ij}^\top \right) - \frac{1}{1 + \sum_{j=1}^{J-1} \eta_{ij}} \sum_{j=1}^{J-1} \eta_{ij} \boldsymbol{\gamma}_{ij} \boldsymbol{\gamma}_{ij}^\top \right]$$

となる．ここで，$\eta_{ij} = \exp\left(\boldsymbol{I}_j^\top X_i^* \boldsymbol{\beta}\right), \boldsymbol{\gamma}_{ij} = \boldsymbol{I}_j^\top X_i^*$ と置いている．

参 考 文 献

Agresti, A. (1992). Analysis of ordinal paired comparison data, *Journal of the Royal Statistical Society*: *Series C* (*Applied Statistics*), **41**(2), 287–297.

Agresti, A. (2010). *Analysis of Ordinal Categorical Data* (*Second Edition*), John Wiley & Sons, New York.

Alaíz, C. M., Barbero, A. and Dorronsoro, J. R. (2013). Group fused lasso, *International Conference on Artificial Neural Networks*, 66–73, Springer, Berlin, Heidelberg.

Ananth, C. V. and Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications, *International Journal of Epidemiology*, **26**(6), 1323–1333.

Anderson, J. A. (1984). Regression and ordered categorical variables, *Journal of the Royal Statistical Society*: *Series B*, **46**(1), 1–22.

Armstrong, B. G. and Sloan, M. (1989). Ordinal regression models for epidemiologic data, *American Journal of Epidemiology*, **129**(1), 191–204.

Bleakley, K. and Vert, J.-P. (2011). The group fused lasso for multiple change-point detection, arXiv preprint, arXiv:1106.4199.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers, *Foundations and Trends® in Machine Learning*, **3**(1), 1–122.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties, *Decision Support Systems*, **47**(4), 547–553.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall/CRC, New York.

Fienberg, S. E. (2007). *The Analysis of Cross-classified Categorical Data*, Springer, New York.

Fienberg, S. E. and Mason, W. M. (1979). Identification and estimation of age-period-cohort models in the analysis of discrete archival data, *Sociological Methodology*, **10**, 1–67.

Goodman, L. A. (1984). *The Analysis of Cross-classified Data Having Ordered Categories*, Harvard University Press, Cambridge, Massachusetts.

川野秀一, 松井秀俊, 廣瀬慧 (2018).『スパース推定法による統計モデリング』, 共立出版, 東京.

McCullagh, P. (1980). Regression models for ordinal data, *Journal of the Royal Statistical Society*: *Series B*, **42**(2), 109–127.

Price, B. S., Geyer, C. J. and Rothman, A. J. (2019). Automatic response category combination in multinomial logistic regression, *Journal of Computational and Graphical Statistics*, **28**(3), 758–766.

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**(2), 461–464.

Snoek, J., Larochelle, H. and Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms, *Advances in Neural Information Processing Systems* **25**, 2951–2959.

Whitehead, A., Omar, R. Z., Higgins, J. P., Savaluny, E., Turner, R. M. and Thompson, S. G. (2001). Meta-analysis of ordinal outcomes using individual patient data, *Statistics in Medicine*, **20**(15), 2243–2260.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society*: *Series B*, **68**(1), 49–67.

Proceedings of the Institute of Statistical Mathematics Vol. 68, No. 2, 287–303 (2020) 303

Fusing Adjacent Classes in an Ordinal Logistic Model via Group Regularization

Mizuho Naganuma[1], Kohei Yoshikawa[2] and Shuichi Kawano[2]

[1]Graduate School of Informatics and Engineering, The University of Electro-Communications; Now at Macromill, Inc.
[2]Graduate School of Informatics and Engineering, The University of Electro-Communications

This paper aims to fuse adjacent classes in an ordinal logistic model in light of the multi-class classification problem. Fusing the classes enables us to easily interpret the constructed model and remove irrelevant classes. Fusion of classes is performed when two adjacent classes have the same posterior probability. To this end, we developed an ordinal logistic model with group regularization for fusing adjacent classes. We established an estimation algorithm based on the alternating direction method of multipliers, and used Monte Carlo simulations and real data analysis to investigate the usefulness of our proposed method.

Key words: Adjacent-categories logit model, alternating direction method of multipliers, group lasso, ordinal categorical data.