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

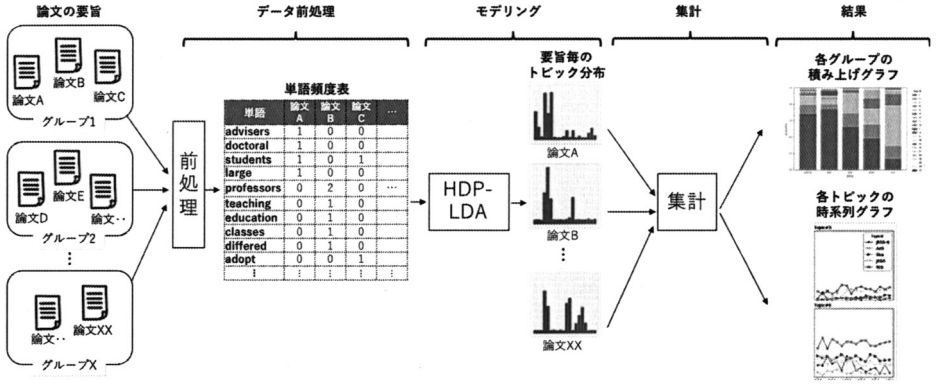

図 1．分析方法の概要．

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

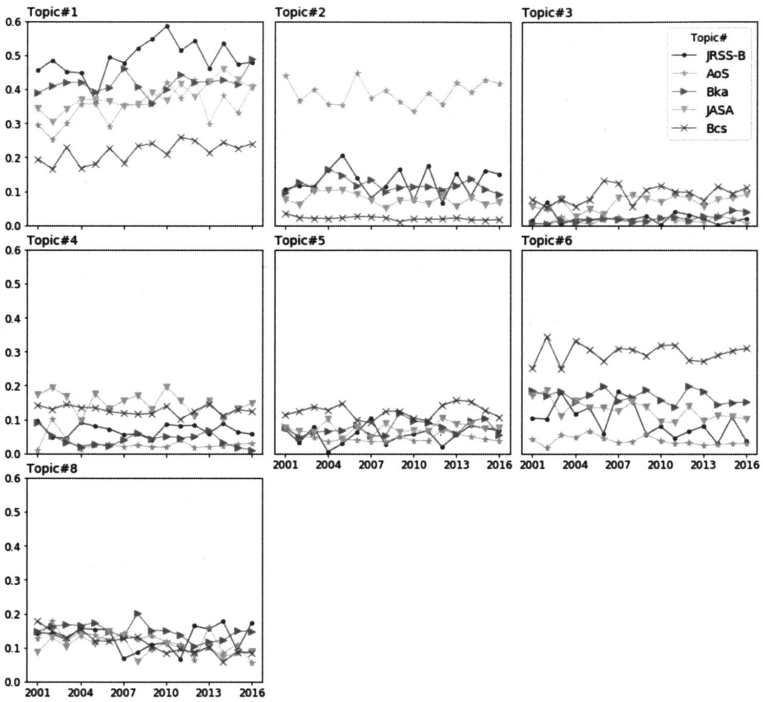

図 2. 統計科学論文誌データのトピック割合の時系列グラフ.

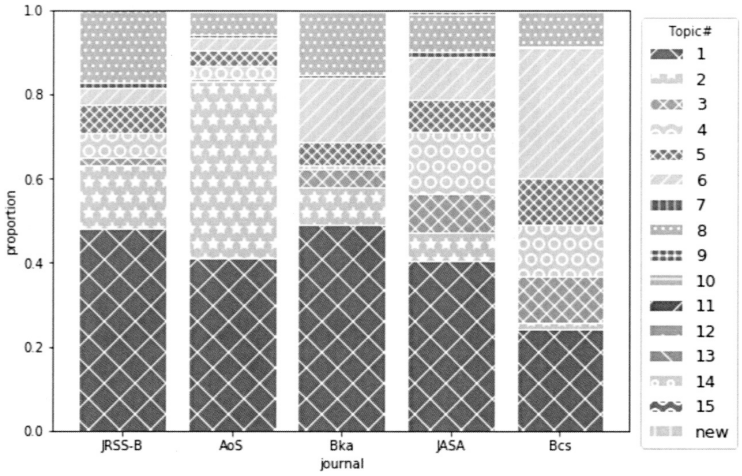

図 3. 統計科学論文誌データの 2016 年度のトピック割合の積み上げグラフ.

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

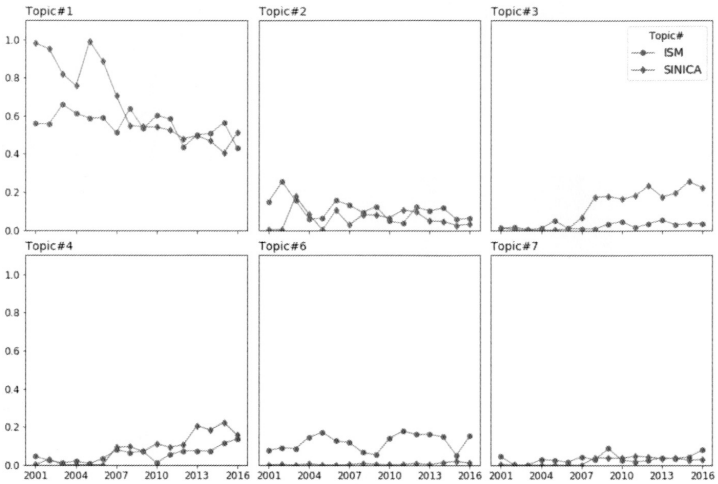

図 4. 研究所データのトピック割合の時系列グラフ.

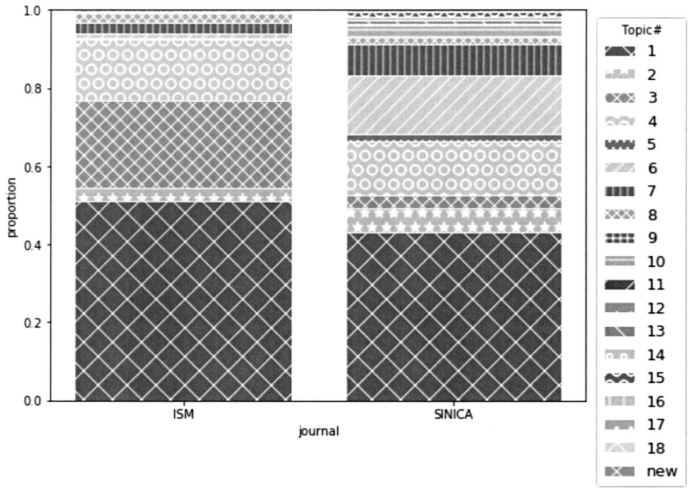

図 5. 研究所データの 2016 年度のトピック割合の積み上げグラフ.

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.