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．

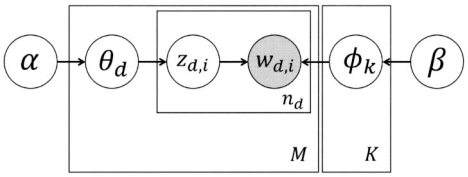

図 1. LDA のグラフィカルモデル表現.

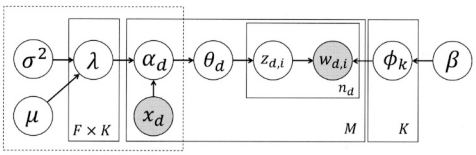

図 2. DMR トピックモデルのグラフィカルモデル表現.

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

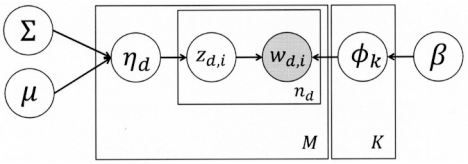

図 3. CTM のグラフィカルモデル表現.

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，

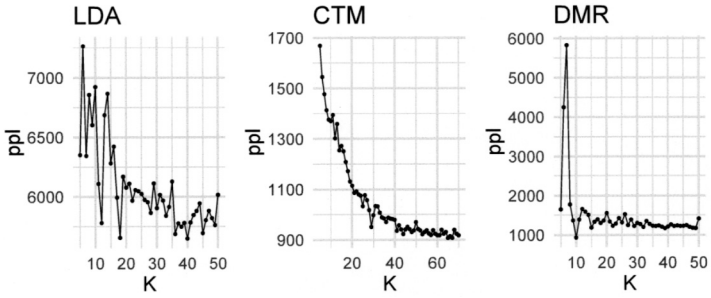

図 4. 各モデルに対する Perplexity.

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

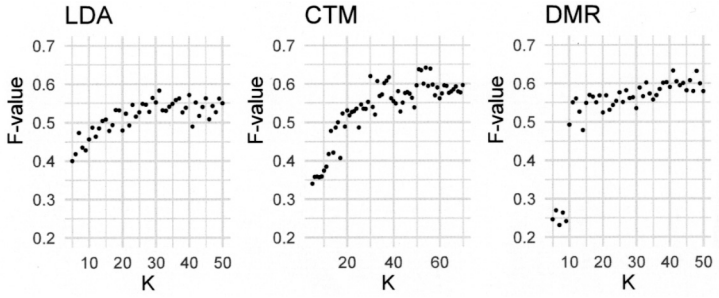

図 5. 各モデルに対する最大の F 値.

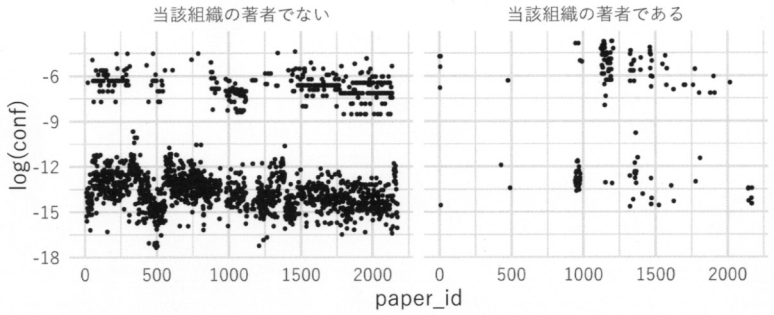

図 6. トピック数を 54 とした CTM による確信度の分布.

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

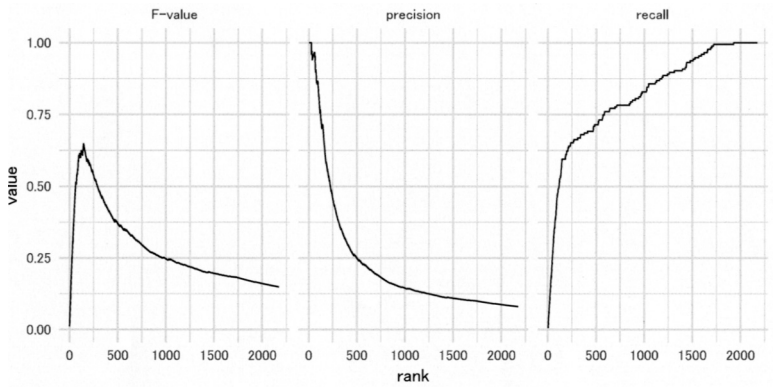

図 7. トピック数を 54 とした CTM による F 値，適合率，再現率.

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.