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

桂井麻里衣[†,††]　大向　一輝[††]　武田　英明[††]

† 日本学術振興会特別研究員
†† 国立情報学研究所　〒101–8430 東京都千代田区一ツ橋 2-1-2
E-mail: †{katsurai,i2k,takeda}@nii.ac.jp

あらまし　本文では，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手法を提案する．提案手法では，代表的な確率的トピックモデルである Latent Dirichlet Allocation (LDA) を用いて学術論文のテキスト情報の次元を削減し，研究者の専門分野をモデル化する．具体的には，各研究者が所有する学術論文のアブストラクトをトピック分布で表現し，それらの平均を研究者の特徴ベクトルとする．また本文では，学術論文データベース整備の高効率化を目的とし，提案手法による著者同定問題への応用について検討する．与えられた学術論文の著者に同姓同名の研究者が存在する場合，研究者の特徴ベクトルを用いてその著者を同定する．CiNii Articles から収集されたデータを用いた実験を行った結果，提案手法により推定したトピック情報を用いることで，同姓同名研究者の論文を正解率 92.99% で同定できることを示した．

キーワード　学術論文，研究者分析，　著者同定，　トピックモデル応用，LDA

1. はじめに

学術論文データベースから主要な技術や研究グループを発見・追跡することは，科学技術の動向の把握や研究戦略の決定，共同研究活動の推進に有用であり，これまでに多くの手法が提案されている [1–6]．従来研究の代表的なアプローチとして，論文の著者名や参考文献などのメタ情報を参照して共著関係や引用関係を表すグラフを構築し，ネットワーク分析手法を適用する方法が挙げられる [1–3]．この方法は，研究者間のソーシャルな関係性を示すことができるものの，研究内容の類似性による潜在的なつながりを見出すことが困難である．一方，学術論文のキーワードやアブストラクトなどのテキストコンテンツに着目することで，各研究者の専門とする研究トピックを明らかにする手法が近年提案されている [5]．これらの手法では，学術論文中で使用される膨大な数の単語から各著者を特徴付ける単語集合を選択するという目的の下，各単語をトピック分布で表現するモデル（以降，トピックモデル）の有効性が示されてきた．トピックモデルを用いることで，データベース内の研究トピックが可視化されると同時に，高次元空間上のテキストコンテンツを低次元空間で表現することが可能となる．

従来研究の多くは，ある限られた研究分野での解析結果を報告している．例として，文献 [3,4] は自然言語処理の分野で出版された学術論文のみを対象としており，文献 [5] は情報科学の分野で顕著な研究者のみを解析している．これに対し，工学から医学まであらゆる研究分野を含む学術論文データベースにおいて高精度に研究者のトピックを推定することができれば，学際的な研究の推進や異分野での共同研究の推薦に貢献できると考えられる．

そこで本稿では，国内最大規模の学術論文情報データベース CiNii Articles(注1) を用いて研究者のトピック情報を推定する．CiNii Articles は国立情報学研究所が 2005 年から運営する学術論文検索・提供サービスであり，全ての論文に書誌 ID が付与されている [7]．この CiNii Articles に収録されている学術論文のアブストラクトを利用して，各研究者が専門とするトピックを推定する．提案手法では，アブストラクトのテキスト情報を低次元空間で表現するために，代表的なトピックモデルである Latent Dirichlet Allocation (LDA) を利用する．まず，データベース中の学術論文の部分集合をトレーニングデータとみなし，LDA により予め研究トピックの抽出を行う．次に，全ての学術論文をトピック分布で表し，各著者に対し平均を算出することで，研究者のトピックを表すベクトルを得る．

提案手法により得られた研究者のトピック情報には様々な応用先が考えられるが，本稿ではその一例として，学術論文の著者同定問題について検討する．ここで著者同定とは，与えられた論文の著者名に複数の同姓同名研究者が該当する場合，正解となる研究者を選出するタスクを指す．実際に，CiNii Articles をはじめとする様々な学術論文データベースの整備において，同姓同名研究者の存在は大きな問題の一つである [8]．提案手法では，新たに与えられた学術論文の著者に同姓同名の研究者が存在する場合，学術論文と著者候補の研究トピックを比較することで著者を同定する．

本研究による主な貢献は以下の通りである．

- 特定の研究分野を対象として学術論文データベースを解析する従来研究とは異なり，あらゆる研究分野を網羅的に含めて研究者のトピック情報を推定する．
- 学術論文の著者同定問題に着目し，研究者のトピック推

(注1)：http://ci.nii.ac.jp/

定の応用可能性について論じる．

本文の構成は以下の通りである．まず，2. において本研究の関連研究を紹介する．3. では，学術論文データベースにおける研究者のトピック推定手法を提案し，4. では提案手法による著者同定への応用について述べる．5. では，実験を行い，提案手法の有効性を評価する．最後に，6. において，本文のまとめと今後の方向性について検討する．

2. 関 連 研 究

2.1 学術文献データベースからの知識抽出

学術論文データベースからの知識抽出は従来より盛んに研究されており，共著関係や引用関係のネットワーク分析を行うことで，学術コミュニティの中で重要な論文や研究者を効果的に発見できることが示されてきた．White，Griffith ら [1] は，ある二人の研究者の論文が他の論文の参考文献で共起するほどその研究者間には強い関係が存在するという共引用関係に着目した．Börner ら [2] は共著関係を論文数による重み付きグラフで表し，ネットワークの中心性を算出して影響力の強い著者を発見した．同様に，Radev ら [3] は自然言語処理に関する著名な国際会議の論文をデータベース化し，ネットワーク分析から得られる指標に基づき論文や著者のランク付けを行った．

一方，データベース内に存在する著者の研究内容の特定や主流研究テーマの変遷の可視化を目的とし，近年では学術論文の内容をトピックで表現する手法が種々提案されている．Hall ら [4] は LDA を用いて主要国際会議のトピック分布を算出し，それらの間の類似度を算出することで国際会議の類似性を分析した．さらに，トピック分布のエントロピーを算出することで，各国際会議が対象とする研究テーマの広さを分析した．研究内容に基づく研究者間類似度の算出を目的としてトピックモデルを用いた手法は，2012 年に Lu，Wolfram ら [5] らによって提案されており，これが本研究と最も関連する従来研究であるといえる．具体的には，LDA に著者情報を導入して拡張した Author Topic Model [9] をデータベースに適用し，学術論文の著者をトピックに関する多項分布で表す．しかしながら，文献 [5] では解析対象とする研究者を 50 名に限定しており，その研究分野は情報科学のみである．それに対し，本稿では，情報科学や工学，教育学，医学などを網羅的に含む大規模学術データベースにおいてトピックモデルを適用する．

学術論文データベースにおける研究トピックの変遷を可視化するために，論文の出版日時を変数として導入したトピックモデルが利用されている．例として，Gerrish ら [6] は Dynamic Topic Model [10] をデータベースに適用し，トピックの変化を利用して各論文の影響力を表す指標を定義した．本研究のように全研究分野を含むデータベースを対象とする場合においても，研究者のトピック推定時にデータベース内のトピックの変遷を考慮すべきかどうかは，今後検討する必要がある．

2.2 学術論文の著者同定に関する研究

学術論文の著者同定は，論文データベースの整備や解析において必要不可欠な技術である．特に，本研究や文献 [5] のように研究者をトピックで表現したい場合，研究者と学術論文が正確に対応付けられている必要がある．このような研究者に関する情報を整理して提供するためには，氏名ではなく識別子 (ID) を用いて予め研究者を区別する必要がある [11]．例として，CiNii Articles では，国内の研究者の名前典拠である研究者リゾルバー [12] で定義された ID に論文を紐付けている．類似したサービスには，Thomson Reuters の ResearcherID(注2) や，Elsevier の Author Identifier(注3) が挙げられる．さらに，これら複数のデータベースと連携して情報流通を行うために，Open Researcher and Contributor ID (ORCID) [13] という組織が世界中の研究者に対し著者 ID を発行しており，登録者数が 2014 年 11 月に 100 万人を突破したことが報告されている．

ウェブ上の研究者 ID に学術論文を紐付ける場合，論文著者の同姓同名研究者の存在が問題となり，依然としてデータベース管理者やユーザによる手動の処理が必要である．Thomson Reuters のデータベース Web of Science や Elsevier の Scopus には独自の著者同定システムが導入されているが，このシステムのみでは高精度な同定は困難であるとの報告がなされている [14]．Strotmann ら [8] は，ネットワーク分析を行う際に，イニシャルの一致性や論文の共著者情報を利用して著者の曖昧性を解消する手法を提案した．一方，研究者の所属は変わる可能性があるほか，単著での論文発表も予想しうる．既存の著者同定の性能を向上させるためには，所属や共著者などのメタ情報のみならず，学術論文の内容を相補的に利用する必要があると考えられる．そこで本研究では，提案手法の一つの応用先として，学術論文の内容と著者候補の研究トピックを比較することで著者同定を行う．

3. 学術論文データベースにおける研究者のトピック推定

本章では，学術論文データベースにおける研究者のトピック推定手法を提案する．提案手法の概要を図 1 に示す．まず，図 1 (a) に示すように，トレーニングデータセットを用いて LDA のパラメータを算出し，研究トピックを表す単語分布を推定する（3.1）．次に，図 1 (b) に示すように，各研究者に紐付いた学術論文のトピック分布を用いて研究者をモデル化する（3.2）．以降の節では，それぞれの具体的な内容について説明する．

3.1 研究トピックを表す単語分布の推定

本節では，データベース中の研究トピックを表す単語分布の推定方法について説明する．まず，データベースから研究分野を網羅的に含む N 本の学術論文を収集し，これをトピック抽出のためのトレーニングセット D とする．論文 $d_i \in D$ のテキスト情報に対し形態素解析を適用して得られた単語セットを $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \cdots, x_{iW_i}]$（$W_i$ は d_i の単語の総数）とし，データセット中のボキャブラリを $V = [w_1, w_2, \cdots, w_W]$（$W$ はボキャブラリのサイズ）で表す．提案手法では，データベース中の

(注2)：http://www.researcherid.com/
(注3)：http://www.elsevier.com/online-tools/scopus

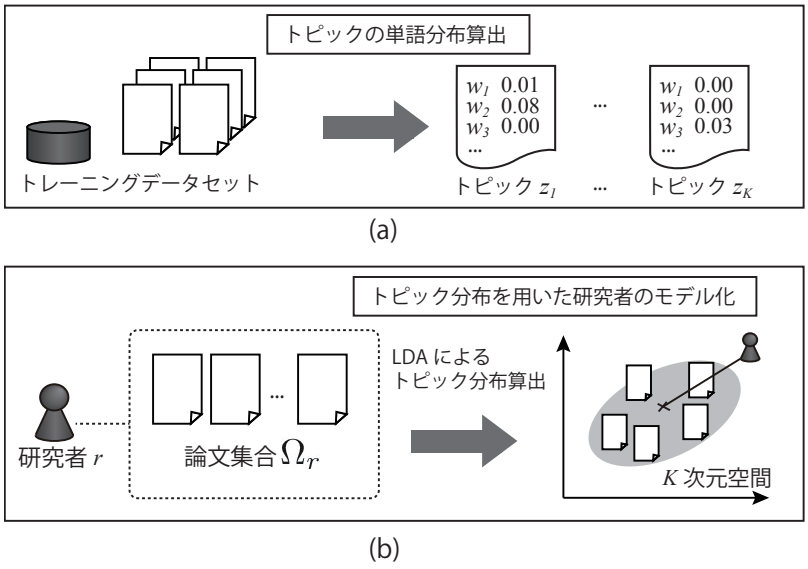

図 1　提案手法による研究者のモデル化の概要.

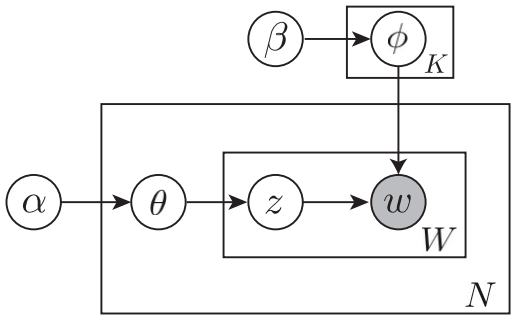

図 2　LDA のグラフィカルモデル.

トピックの情報と学術論文のトピックの混合比を求めるために，LDA を用いる．LDA のグラフィカルモデルを図 2 に示す．今，データベース中に存在する K 個のトピック $\boldsymbol{z} = [z_1, z_2, \cdots, z_K]$ を考えたとき，各トピックをボキャブラリについての多項分布で表し，かつそれらがディリクレ分布から生成されていると仮定する．図 2 のモデルに基づき，データセット D 中の学術論文の生成プロセスは次のように表される．

1. 文書 d_i に対し，$\boldsymbol{\theta}_i \sim Dir(\alpha)$.
2. $k = 1, 2, \cdots, K$ 番目のトピックに対し，$\boldsymbol{\phi}_k \sim Dir(\beta)$.
2. 文書中の単語 $x_{ij} \in \boldsymbol{x}_i$ に対し，トピック z_{ij} と単語 x_{ij} を次のように選択する．

- $z_{ij} \sim Mult(\boldsymbol{\theta}_i)$
- $x_{ij} \sim Mult(\boldsymbol{\phi}_{z_{ij}})$

上式において，文書のトピックに対する多項分布のパラメータ $\boldsymbol{\theta}_i$ とトピックの単語に対する多項分布のパラメータ $\boldsymbol{\phi}_k$ はそれぞれ超パラメータが α, β となるディリクレ事前分布をもつ．$\boldsymbol{\phi}_k$ は k 番目のトピックにおいてどのような単語が重要かを表し，$\boldsymbol{\theta}_i$ は学術論文 d_i にどのようなトピックが含まれるかを表すと解釈できる．本稿では，これらのパラメータを Collapsed ギブスサンプリング [15] によって算出する．具体的には，単語 w がトピック z_k に割り当てられた回数を m_{wk}，文書 d_i 中の単語がトピック z_k に割り当てられた回数を n_{ik} とすると，変数 z_{ij} を除く全ての状態が与えられたときの z_{ij} の事後確率は次式で算出される．

$$P(z_{ij} = k | \boldsymbol{z}^{\neg ij}, \boldsymbol{x}, \alpha, \beta) = \frac{1}{Z} a_{ik} b_{wk} \tag{1}$$

上式において，

$$a_{ik} = n_{ik}^{\neg ij} + \alpha \tag{2}$$

$$b_{wk} = \frac{m_{wk}^{\neg ij} + \beta}{m_{\cdot k}^{\neg ij} + W\beta} \tag{3}$$

$$Z = \sum_k a_{ik} b_{wk} \tag{4}$$

であり，添字 $\neg ij$ は対応するデータ $x_{i,j}$ を式中の総和の算出に利用していないことを表す．ギブスサンプリングを繰り返すことで，式 (1) により論文 d_i 中の単語 x_{ij} に対する z_{ij} が推定され，この値に基づき式中の総和を更新する．最終的に，$\boldsymbol{z}$ が与えられたときの $\boldsymbol{\phi}_k, \boldsymbol{\theta}_i$ の予測分布は次式で推定される．

$$\hat{\phi}_{wk} = \frac{m_{wk} + \beta}{m_{\cdot k} + W\beta} \tag{5}$$

$$\hat{\theta}_{ik} = \frac{n_{ik} + \alpha}{n_{i\cdot} + K\alpha} \tag{6}$$

以上のように LDA を用いて算出されたトピック分布を学術論文の低次元表現と考え，研究者のモデル化を行う．

3.2　学術論文のテキスト情報を用いた研究者のトピック推定

本節では，学術論文のテキスト情報を用いた研究者のトピック推定手法を提案する．モデル化対象とするデータベース中の研究者を r で表し，研究者 r が著者となる学術論文集合を Ω_r とする．研究者 r の各学術論文 $d_i \in \Omega_r$ より，ボキャブラリ V に含まれる単語セットを抽出する．LDA により推定された学術論文 d_i のトピック混合比 $\boldsymbol{\theta}_i$ に基づき，研究者 r のトピックを表すベクトルを次式により算出する．

$$\boldsymbol{m}_r = \frac{1}{|\Omega_r|} \sum_{d_i \in \Omega_r} \boldsymbol{\theta}_i \tag{7}$$

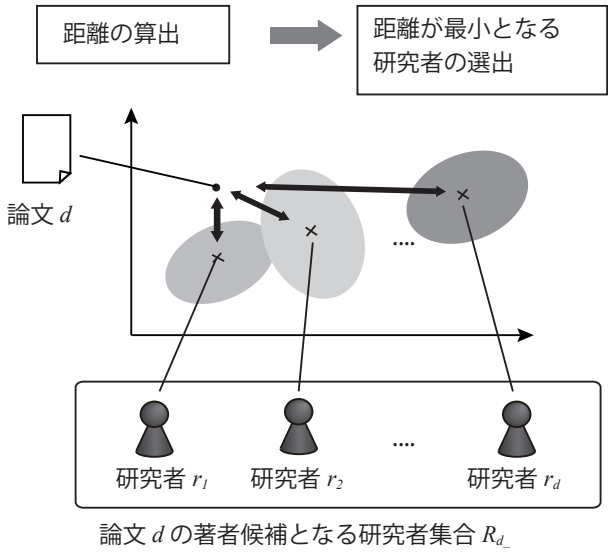

図 3　提案手法による学術論文の著者同定.

ここで，$|\Omega|$ は集合 Ω の要素数を表す．上式のように，研究者 r がこれまでに発表した学術論文のテキスト情報を用いることで，研究者 r の研究トピックを表すことが可能となる．

研究者をモデル化することで，様々な応用先が考えられる．例として，研究者 r_1 と r_2 のトピック類似度を算出する場合には，ベクトル $\boldsymbol{m}_{r_1}$ および $\boldsymbol{m}_{r_2}$ 間の類似度を算出すればよい．

提案した研究者のトピック推定手法は，研究者 r とその著作となる学術論文集合がデータベース内で紐付けられていることを想定している．一方，2. で述べたように，データベース中には多数の同姓同名研究者が存在するため，著者候補者から正しい著者の選定が必要となる．そこで次章では，提案手法で推定された研究者のトピック情報に基づく著者同定について説明する．

4. 研究者のトピック情報に基づく学術論文の著者同定

本章では，提案手法の応用として，研究者のトピック情報に基づく学術論文の著者同定について説明する．提案手法による著者同定の概要を図 3 に示す．図に示すように，データベースに新たな学術論文が与えられたとき，論文が取り扱う研究トピックと著者候補者の専門とする研究トピックを比較することで，研究者の選出を行う．これにより，新たに論文が研究者に紐付けられ，データベースが更新される．

具体的には，新たに与えられた学術論文を d_{new} とし，d_{new} の著者名セットを $A_{d_{new}}$ とする．また，データベース中に存在する氏名 $a \in A_{d_{new}}$ と同姓同名の研究者セットを R_a とする．$|R_a| > 1$ となる著者名 $a \in A_d$ が存在する場合，次式により研究者 r^* を選出する．

$$r^* = \arg\max_{r \in R_a} sim(\boldsymbol{m}_r || \boldsymbol{\theta}_{new}) \tag{8}$$

上式において，$\boldsymbol{\theta}_{new}$ は LDA により推定された d_{new} のトピック分布であり，$sim(\boldsymbol{m}||\boldsymbol{\theta})$ はベクトル $\boldsymbol{m}$，$\boldsymbol{\theta}$ 間の類似度を表す．

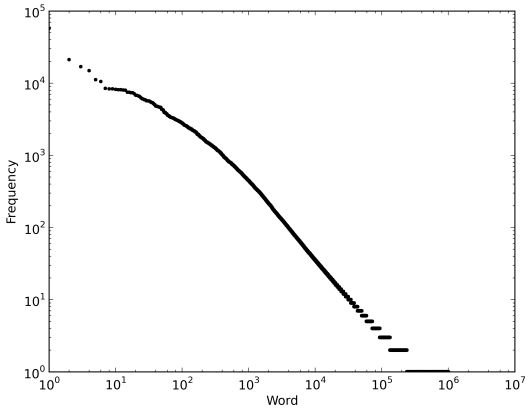

図 4　トレーニングデータセットにおける単語の頻度と順位の散布図.

表 1　トレーニングデータセットにおける最頻出単語.

結果，研究，検討，可能，明らか，必要，方法，実験，目的，報告，提案，場合，影響，問題，変化，対象，比較，評価，関係，利用，本稿，以上，システム，調査，重要，測定

表 2　トレーニングデータセットにおける出現回数が 5 未満の単語の例．これらの単語の一部は名詞のバイグラムによって置き換えられる.

順応的管理プログラム， 電子顕微鏡的解析，シュレディンガー方程式バンド，アルダーソン， 高性能プローブ，ボトルネックネットワーク， 電気抵抗変化型 DNA 検出法，昼寝時，移動動作時，木造密集住宅地，静止画像圧縮法，リンケージ同定手法，歴史地区指定，アイスアルジー原生生物，未知システム係数ベクトル，インヒビン遺伝子，喘息期間有症率，購入余裕，新設施設投資

なお，式 (8) に閾値を導入し，与えられた論文の著者がデータベースに含まれない新規著者であるかどうかを判断することも可能となるが，本稿では取り扱わず，今後の課題とする．

提案手法により学術論文 d_{new} の著者を同定することで，データベースの整備を行うと同時に研究者のトピック情報を更新することが可能となる．更新は次式で表される．

$$\boldsymbol{m}_r \leftarrow \frac{1}{1+|\Omega_r|}\left(\boldsymbol{\theta}_{new} + |\Omega_r|\boldsymbol{m}_r\right) \tag{9}$$

$$\Omega_r \leftarrow \Omega_r \cup \{d_i\} \tag{10}$$

以上のように，LDA によって得られたトピックの単語分布と，学術論文のトピック分布に基づき，学術論文の著者同定および研究者のトピック情報の更新が可能となる．

5. 実　　験

本章では，提案手法の有効性を示すために実験を行う．まず，5.1 において実験で用いるデータセットについて述べる．次に，5.2 において LDA により推定されたトピックの単語分布の内容を報告する．最後に，5.3 において，著者同定への応用による評価を行う．

5.1 データセット

本節では，実験で利用したデータセットの詳細を述べる．ま

表 3　トレーニングデータセットの詳細.

学術論文数	59,711
研究者数	31,399
単語の総数	61,400
1 論文あたりの単語数の平均	28

表 4　著者同定の評価に用いるデータセット.

学術論文数	5,267
研究者数	345
同姓同名が存在する氏名数	166
1 氏名あたりの同姓同名研究者の数	2.078

ず，科研費データベース KAKEN(注4) に収録されている全研究者のうち，CiNii Articles 上に論文のアブストラクトが存在する研究者 ID を抽出し，ID を羅列したリスト（以降，研究者リストと呼ぶ）を作成した．この研究者一名につき最大 5 本の和文アブストラクトを収集することで，合計 59,771 本の学術論文を得た．得られた学術論文集合は国内の研究分野を網羅的に含むと考えられ，これを 3.1 で述べた LDA のトレーニングセットとして利用する．

次に，各学術論文のアブストラクトから単語セットを抽出する方法について述べる．まず日本語 Wikipedia のエントリを辞書に登録し，MeCab(注5) により形態素解析を行った．今回の実験ではすべての名詞のユニグラムをテキスト特徴として抽出した．さらに，学術用語の中には Wikipedia に登録されていないものも存在するため，名詞が連続する場合には全てのバイグラムも抽出した．以上の方法でトレーニングデータセットから抽出された単語の頻度と，頻度による順位の関係を図 4 に示す．ここで，縦軸は単語の頻度，横軸は単語の頻度の順位を表し，いずれも対数スケールである．この図のうち，頻度が上位となる単語を表 1 に示す.「研究」「結果」「検討」などの単語は学術論文を特徴付けるうえで効果が低いと考えられるため，実験ではこれらの最頻出単語をストップワードとして扱うこととする．さらに，データセット内で出現回数が 5 未満となる低頻度の単語についてもボキャブラリから除外する．除外した低頻度の単語の例を表 2 に示す．なお，低頻度かつ Wikipedia に登録されていない単語については，語の一部のバイグラムがボキャブラリに追加されている可能性がある．例として，表中の「静止画像圧縮法」という単語は,「静止画像」および「画像圧縮」という単語によって置き換えられる．以上のように構築された LDA のトレーニングデータセットの詳細を表 3 に示す．

続いて，研究者リストから全ての同姓同名研究者とその氏名を抽出し，それらが著者となる学術論文のうち和文アブストラクトを含むものを全て収集した．得られたデータセットを表 4 に示す．5.3 では，このように多数の同姓同名研究者によって構成されるデータセットを用いて，著者同定の性能を評価する．

5.2　LDA によるトピック推定結果

まず，5.1 で用意したトレーニングデータセットを用いて，3.1 で説明したようにトピックの単語分布を算出した．今回の実験では LDA のトピック数を $K = 500$ とし，LDA の超パラメータを $\alpha = \frac{50}{K}, \beta = 0.01$ と設定した．抽出されたトピックの例を表 5 に示す．表 5 では，科学研究費助成事業によって定められた研究分野表を参考にし，著者らが手動で付与したトピックの解釈を太字で示す．この表から，LDA を用いることで，多様な分野の研究者から収集された学術論文から各研究分野を表すトピックが効果的に抽出されているといえる．

一方，抽出されたトピックは，専門の広さが異なることも確認できる．例として，トピック 97 や 213 は，ある特定の分野を特徴付ける単語から構成されているが，トピック 129 は工学・情報科学の分野で広く用いられる単語から構成されている．今回はこうしたトピックの階層性の存在を考慮せずに研究者のモデル化を行う．また，トピック 446 の画像処理のように，様々な研究分野において利用される手法のトピックも存在することから，これらのトピックを用いて研究者のモデル化を行うことで，異分野を横断した類似研究者の検索が行える可能性があるといえる．

表 6　著者同定の実験結果.

正しく著者を同定した論文の数	4,898
誤って著者を同定した論文の数	369
正解率	92.99%

5.3　著者同定の評価

本節では，提案手法による学術論文の著者同定の実験を行う．まず実験方法について述べる．実験では，各研究者 ID に対し，その出版物の中で最も古い論文が最低一本は確実に紐付いている状態を想定する．データセット中の著者名 a に対する同姓同名者集合 R_a に対し，著者名 a を持つ学術論文を出版年月日により昇順に並べたとき，与えられた論文 d の著者を式 (8) により同定し，これを一回のテストとみなす．次に，結果の正解・不正解に関わらず，論文 d を用いて正しい著者のトピック情報を式 (9) により更新する．以上のテストを全ての同姓同名者集合に対し行い，次式の正解率を算出した．

$$\text{正解率} = \frac{\text{正しく著者を同定した論文の数}}{\text{著者同定の対象となる論文の数}} \times 100 \quad [\%] \tag{11}$$

得られた結果を表 6 に示す．結果は正解率 92.99% という高い数値を示した．したがって，学術論文の内容と研究者のトピックを比較することで，高精度に著者を同定できることが示された．

ここで，著者同定を誤った学術論文の例を表 7 に示す．表 7 の例 1 の著者候補となる同姓同名研究者は，知覚情報処理を専門とする研究者 A，教育工学を専門とする研究者 B，および農業環境工学を専門とする研究者 C の三名であった．正しい著者は C であるが，手法によって A と推定したため，誤同定となった．これはアブストラクト中の「位置検出」「アルゴリズム」「検出率」などの単語の表すトピックが研究者 A のトピックに類似していたことが原因であると考えられる．

例 2 では，形成外科学を専門とする研究者 D，基礎看護学を

(注4)：https://kaken.nii.ac.jp/
(注5)：https://code.google.com/p/mecab/

表 5　LDA によって推定された研究トピックの例．トピックの解釈を太字で示す．

トピック 7 **反応工学**		トピック 13 **人体病理学**		トピック 123 **教育工学**		トピック 133 **教育学**		トピック 160 **神経科学**		トピック 251 **自然言語処理**	
単語	確率	単語	確率	単語	確率	単語	確率	単語	確率	単語	確率
生成	0.0743	腺癌	0.0281	大学	0.0879	教師	0.1267	神経細胞	0.0365	単語	0.0406
反応	0.0422	腫瘍細胞	0.0213	高等教育	0.0454	授業	0.0492	神経回路	0.0315	語	0.0338
酸化	0.0389	免疫組織化学的	0.0192	学生	0.0369	指導	0.0426	応答	0.0288	文	0.0301
生成物	0.0308	組織	0.0183	教育	0.0313	生徒	0.0182	情報伝達	0.0277	文章	0.0225
分解	0.0236	間質	0.0171	教育機関	0.0273	実践	0.0179	軸索	0.0208	意味的	0.0223
化学反応	0.0230	染色性	0.0167	大学教育	0.0216	学び	0.0179	ニューロン	0.0193	辞書	0.0208
熱分解	0.0194	学的	0.0165	本学	0.0187	子ども	0.0156	回路網	0.0179	自然言語	0.0208
化学種	0.0170	陽性細胞	0.0165	教養教育	0.0118	実践的	0.0138	神経系	0.0177	名詞句	0.0176
進行	0.0163	線維化	0.0152	遠隔教育	0.0115	学校	0.0136	生理学的	0.0164	抽出	0.0163
生成量	0.0153	細胞	0.0149	教育システム	0.0091	分析	0.0131	可塑性	0.0153	使用頻度	0.0131

トピック 25 **システム**		トピック 47 **インタラクション**		トピック 83 **遺伝学**		トピック 243 **心理学**		トピック 183 **消化器外科学**		トピック 375 **経済政策**	
単語	確率	単語	確率	単語	確率	単語	確率	単語	確率	単語	確率
装置	0.1316	コミュニケーション	0.0855	遺伝子	0.0444	社会的	0.2679	視鏡	0.0407	中国	0.0383
開発	0.1085	エージェント	0.0537	塩基配列	0.0393	文化的	0.0529	内視	0.0406	経済発展	0.0237
センサ	0.0421	自律的	0.0457	解析	0.0229	包括的	0.0475	腹腔鏡	0.0398	国際化	0.0219
試作	0.0380	対話	0.0415	配列	0.0205	規則性	0.0362	鏡下	0.0251	日本	0.0216
製作	0.0372	発話	0.0408	アミノ酸配列	0.0199	的側面	0.0256	手術	0.0211	グローバル化	0.0213
実験装置	0.0360	会話	0.0358	欠失	0.0196	認知的	0.0235	手術時間	0.0182	発展	0.0150
小型	0.0239	相手	0.0260	同性	0.0191	的要因	0.0215	鏡的	0.0154	経済	0.0132
応用	0.0221	共有	0.0193	コード	0.0190	社会	0.0179	術者	0.0153	先進国	0.0130
使用	0.0212	実現	0.0180	決定	0.0188	的背景	0.0178	出血量	0.0141	経済成長	0.0129
感度	0.0202	人間	0.0133	検出	0.0155	認知	0.0174	施行	0.0139	急速	0.0126

トピック 31 **森林科学**		トピック 138 **水産学**		トピック 129 **統計科学**		トピック 341 **都市計画**		トピック 213 **電子デバイス**		トピック 391 **言語学**	
単語	確率	単語	確率	単語	確率	単語	確率	単語	確率	単語	確率
本種	0.0544	海域	0.0240	数値計算	0.0852	都市	0.0417	屈折率	0.0739	言語	0.0503
数種	0.0434	底質	0.0171	解析的	0.0393	居住者	0.0294	導波路	0.0673	日本語	0.0499
成虫	0.0271	底層	0.0150	数値的	0.0390	都市化	0.0222	光ファイバ	0.0264	英語	0.0416
幼虫	0.0261	栄養塩	0.0140	適用	0.0288	住宅	0.0203	光	0.0168	外国語	0.0192
採集	0.0234	海水中	0.0131	解	0.0273	建築物	0.0185	作製	0.0158	中国語	0.0178
種	0.0200	植物プランクトン	0.0124	行列	0.0213	住宅地	0.0154	構成	0.0121	英語教育	0.0177
個体	0.0183	漁獲量	0.0123	数値実験	0.0199	居住地	0.0133	導波	0.0118	分析	0.0174
寄生	0.0133	流入	0.0121	計算	0.0193	中心市街地	0.0126	波長	0.0114	語彙	0.0155
発育	0.0131	負荷量	0.0113	解析	0.0192	居住環境	0.0099	構造	0.0108	格助詞	0.0147
齢幼虫	0.0113	水塊	0.0107	離散化	0.0175	場所	0.0098	光スイッチ	0.0108	意味	0.0131

トピック 97 **通信・ネットワーク工学**		トピック 147 **心理学**		トピック 186 **環境学**		トピック **社会学**		トピック 266 **スポーツ科学**		トピック 446 **画像処理**	
単語	確率	単語	確率	単語	確率	単語	確率	単語	確率	単語	確率
符号化	0.1525	心理的	0.1189	濃度	0.1016	制度	0.0264	運動	0.1453	画像	0.1324
符号	0.0348	ストレス	0.0606	大気中	0.0681	自治体	0.0195	健康	0.0383	画像処理	0.0506
情報源	0.0252	不安	0.0432	co	0.0346	事業	0.0169	運動郡	0.0361	画像データ	0.0261
適用	0.0215	生理的	0.0363	一度	0.0296	課題	0.0148	身体活動	0.0342	原画像	0.0228
符号語	0.0213	ストレス反応	0.0264	測定	0.0229	政策	0.0146	スポーツ	0.0277	撮影	0.0201
量子化	0.0207	的ストレス	0.0212	発生源	0.0179	国	0.0144	運動量	0.0251	手法	0.0173
通信路	0.0168	心理状態	0.0212	大気汚染物質	0.0134	市町村	0.0133	運動能力	0.0234	解像度	0.0139
復号法	0.0151	身体的	0.0212	ガス	0.0122	万人	0.0119	運動習慣	0.0220	画素	0.0139
最適	0.0147	示唆	0.0195	大気汚染	0.0120	政府	0.0112	身体運動	0.0218	劣化	0.0128
漸近的	0.0147	精神的	0.0187	放出	0.0119	推進	0.0110	体力	0.0204	復元	0.0128

表 7　著者同定を誤った論文アブストラクトの例．実際の著者である研究者の専門分野と，手法により誤って選出した研究者の専門分野を示す．

	アブストラクト	正解の研究者の専門分野	誤って選出した研究者の専門分野
例 1	本研究では, イチゴ収穫用として, フックによる対象果実と隣接果実の分離, 誘導および果柄の切断を行う, フック式エンドエフェクタと, 果柄の位置検出アルゴリズムを開発し, 収穫率, 検出率および許容誤差について明らかにした．(http://ci.nii.ac.jp/naid/110002506140)	農業環境工学	知覚情報処理
例 2	骨格や臓器の 3 次元形状の統計モデルは，基礎および臨床医学において大きな期待を集めているが，形成再生外科領域で重要である頭蓋顎顔面骨は，形状の複雑さなどから着手が遅れている．本論文では，頭蓋顎顔面骨データベース構築を目的として，複雑な形状の安定かつ効率的なメッシュ変形手法を提案する．この手法を用いてテンプレートメッシュを変形し，サンプルの計測点の間で密な対応付けを行う．提案手法は，エッジベースの測地距離および Voronoi 分割に基づきメッシュの変形を行う．これにより，位相的に複雑な形状への適用を可能とし，他の操作上の要求も満足させる．CT 画像を用いて実験を行ったところ，提案手法の実用可能性が示された．(http://ci.nii.ac.jp/naid/110006595169)	形成外科学	基礎看護学
例 3	波面合成法による立体音場再生システムを実用化するためには定位実験を実施することが非常に重要である．しかし，定位実験を実施するには多大な時間と労力を要する．本論文では定位実験に要する時間と手間を削減するために，定位実験の結果を予測することができる定位モデルを提案する．定位モデルを実装したところ，定位実験における知覚方向を十分に予測できることが分かった．(http://ci.nii.ac.jp/naid/110006391889/)	知能情報学	ソフトウェア
例 4	本論文は，任意形状に最適なアンテナパターン形成法を提案し，一例として半円形状アレーアンテナに対する種々のアンテナパターンを計算し，その有効性を検討したものである．(http://ci.nii.ac.jp/naid/110003281377)	電力工学	呼吸器内科学

専門とする研究者 E が著者候補者であったが，誤って研究者 E が選出された．この誤同定は両者の研究トピックが類似しているためであると考えられる．同様に，例 3 では，知能情報学を専門とする研究者 F，ソフトウェアを専門とする研究者 G が著者候補者であったが，誤って研究者 G が選出された．これについても，与えられた論文アブストラクトから抽出された単語が少なく，かつ両者の研究トピックが類似していたことが誤同定の原因であったといえる．今後は，同分野内の研究トピックをさらに細分化するためのアルゴリズムの導入についても検討する必要がある．

最後に，例 4 の著者候補となる同姓同名研究者は，電力工学を専門とする研究者 D と呼吸器内科学を専門とする研究者 E の二名であった．正しい著者は D であるが，手法によって E と推定したため，誤同定となった．この原因として，研究者 D のこれまでの論文にアンテナパターンに関する文献が存在せず，その他の電力系の研究との類似性が利用できていないことが挙げられる．この問題を解決するにはトピック間の類似度を考慮した手法が効果的であると考えられる．

なお，提案手法では与えられた学術論文のアブストラクトのみを利用して著者同定を行ったが，著者の所属機関などのメタ情報も同時に利用することでより高精度な同定を行えることが予想できる．

6. まとめと今後の課題

6.1 ま と め

本研究では，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手法を提案し，その著者同定への応用について検討した．提案手法では，論文アブストラクトのテキスト情報を低次元空間で表現するために，トレーニングデータセットに対し代表的なトピックモデルである LDA を適用し，トピックの単語分布を算出した．次に，LDA を用いて各研究者に対応付けられている学術論文集合のトピック分布を算出し，その平均を算出することで研究者のトピックを表すベクトルを得た．CiNii Articles のデータを用いた実験を行った結果，提案手法は多様な研究分野を表すトピックを抽出できていることが目視により確認された．さらに，手法による著者同定の実験では，5,267 本の論文に対し正解率 92.99% を達成した．これにより，与えられた学術論文の研究トピックと著者候補となる研究者のトピック情報を比較することで，著者を同定できることが確認できた．

6.2 今後の課題

本節では，提案手法が抱える課題と今後の解決方針について検討する．まず，研究者のモデル化方法について述べる．提案手法では，研究者のトピック情報を推定するために，学術論文のトピック分布の平均を算出した．このとき，論文の被引用数などによる重み付けや，より複雑な形状の分布推定も効果的となる可能性が高い．したがって，これらのモデル化方法についても実験で比較する必要がある．

また，提案手法で用いるトピックモデルの変数についても検討する必要がある．例として，LDA のグラフィカルモデルに学術論文の論文誌名や学会名を導入することで，より各研究分野のトピックを特徴付けることが可能になると考えられる．さらに，5.3 で述べたように，研究トピックには階層性が見られるため，各研究分野を細分化していくトップダウン的なアプローチとの比較を行う必要がある．また，今回は全ての年代を通して固定の研究トピックを利用したが，文献 [6] のように各年代に特有の研究トピックを抽出することも今後の課題とする．こ

れと関連して，新たな研究用語が追加されたときの辞書登録についても検討する必要がある．例として，「ビッグデータ」や「ディープラーニング」などの用語は近年大きな注目を集めているが，これらの用語がどのようなトピックを持つかを論文から推定することも重要である．今後は，これらの事項を検討し，データベース中のトピックモデリングを改良するとともに，新たな用語の辞書追加手法を提案する予定である．

最後に，提案手法の応用について述べる．今回の著者同定実験では，与えられた学術論文のアブストラクトのテキスト情報のみを利用した．より高精度な著者同定のためには，従来研究と同様，著者の所属などのメタ情報を参照することが効果的であると考えられる．今後は，メタ情報とテキスト情報を相補的に利用した新たな著者同定手法について検討する予定である．なお，研究者のトピック推定は，著者同定のみならず，様々な応用先が考えられる．例として，研究者間のトピック類似度を算出することで，共同研究者の推薦が可能となる．こうした効果的な応用方法についても手法の改善と同時に検討する予定である．

文　献

[1] H. D. White and B. C. Griffith. Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, Vol. 32, No. 3, pp. 163–171, 1981.

[2] K. Börner, L. Dall'Asta, W. Ke, and A. Vespignani. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity*, Vol. 10, No. 4, pp. 57–67, 2005.

[3] D. R. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara. The ACL anthology network corpus. *Language Resources and Evaluation*, Vol. 47, No. 4, pp. 919–944, 2013.

[4] D. Hall, D. Jurafsky, and C. D. Manning. Studying the history of ideas using topic models. In *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 363–371, 2008.

[5] K. Lu and D. Wolfram. Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *Journal of the American Society for Information Science and Technology*, Vol. 63, No. 10, pp. 1973–1986, 2012.

[6] S. Gerrish and D. M. Blei. A language-based approach to measuring scholarly impact. In *Proc. Int. Conf. Machine Learning (ICML)*, pp. 375–382, 2010.

[7] 大向一輝. CiNii Articles のシステムデザインとデータモデル (<特集>データベース構築の今). 情報の科学と技術, Vol. 62, No. 11, pp. 473–477, November 2012.

[8] A. Strotmann, D. Zhao, and T. Bubela. Author name disambiguation for collaboration network analysis and visualization. *Proc. American Society for Information Science and Technology*, Vol. 46, No. 1, pp. 1–20, 2009.

[9] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers. Learning author-topic models from text corpora. *ACM Trans. Information Systems*, Vol. 28, No. 1, pp. 4:1–4:38, January 2010.

[10] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proc. Int. Conf. Machine learning (ICML)*, pp. 113–120, 2006.

[11] M. Enserink. Are you ready to become a number? *Science*, Vol. 323, No. 5922, pp. 1662–1664, 2009.

[12] K. Kurakawa, H. Takeda, M. Takaku, A. Aizawa, R. Shiozaki, S. Morimoto, and H. Uchijima. Researcher Name Resolver: identifier management system for japanese researchers. *International Journal on Digital Libraries*, Vol. 14, No. 1-2, pp. 39–58, 2014.

[13] L. L. Haak, M. Fenner, L. Paglione, E. Pentz, and H. Ratner. ORCID: a system to uniquely identify researchers. *Learned Publishing*, Vol. 25, No. 4, pp. 259–264, October 2012.

[14] L. Tang and J. P. Walsh. Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, Vol. 84, No. 3, pp. 763–784, 2010.

[15] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, Vol. 101, No. suppl 1, pp. 5228–5235, 2004.

[16] X. Wang, A. McCallum, and X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proc. Int. Conf. Data Mining (ICDM)*, pp. 697–702, 2007.