オーサートピックモデルを用いた論文分析による潜在的研究グループの発掘に関する研究

小野 龍太郎*　　冨浦 洋一*　　田中 省作†　　上瀧 恵里子‡
*九州大学統合新領域学府ライブラリーサイエンス専攻
†立命館大学文学部
‡九州大学研究戦略企画室

1　はじめに

学際的な研究の促進において，研究戦略や知財管理を行うマネジメント人材（リサーチ・アドミニストレータ）の存在は重要である．リサーチ・アドミニストレータには学際領域の研究を企画・マネジメントする能力が求められる．本研究では，リサーチ・アドミニストレータが異分野を横断する研究を企画・マネジメントする際に補助となるシステムを開発・提案することを目的とする．本システムはあくまでリサーチ・アドミニストレータの補助を目的としており，共同研究の可能性がある研究者の組の候補とその根拠となるデータを提示し，それをリサーチ・アドミニストレータが評価することを想定している．

2　本研究のアプローチ

本研究では論文著者の興味 (トピック) を推定することで共同研究の可能性を発掘する．そこで，著者に対するトピックの生成確率分布 θ とトピックに対する語の出現確率分布 ϕ を Gibbs Sampling を用いて推定する Author-topic-model[1][2] を用いる．Author-topic-model ではドキュメント中に出現する各単語位置に対し，共著者内の一人の著者，トピック，語彙をモデルに従い確率的に割り当てることでドキュメントを生成する．このとき，ドキュメント d（単語列としては，$w_1 w_2 \cdots w_N$，N は単語数）は，各単語位置 n（$n = 1, 2, \cdots,\ N$）に関して以下を繰り返すことで生成される．

1. ドキュメント d を記した著者群 A^d からランダムに著者 x を選択する．
2. 著者 x が有するトピック多項分布 $\theta^{(x)}$ に基づいて，トピック z を生成する．
3. トピック z が有する単語多項分布 $\phi^{(z)}$ に基づいて単語 w_n を生成する．

Gibbs Sampling によって推定された２つの確率分布 θ と ϕ を用いて本研究では２つの分析方法で実験を行った．

2.1　共通のトピックを用いた方法

分析する著者の間に存在する共通のトピックを軸に共同研究の可能性を発掘する方法で，分析対象である著者間に有意に発生確率の高い共通のトピックと著者がそれぞれ固有に有するトピックが存在すれば共同研究の可能性が高いと考える手法である．この手法は，共同研究というものをそれぞれが持つ専門性 (固有のトピック) を活かしながら共通のテーマ (トピック) で行う研究であると仮定している．

部局が異なる研究者 A，B が共著者になっている論文をすべて削除した論文集合を用いてトピック分析し，上記の方法で研究者 A，B が共同研究の可能性がある研究者の組として抽出されるかという予備的な実験を行った．残念ながら，対象とした研究者 A，B の組は抽出されなかった．しかし，A と共通のトピックを有し，かつ B とも共通のトピックを有する第三の研究者 C が見つかり，(A，C)，(B，C) それぞれが共同研究の可能性がある研究者の組として抽出された．

このことは，単に二人の研究者の組を上記の条件で探索するのではなく，二人の研究者を媒介する第三の研究者まで探索することで，共同研究につながる３名以上から成る研究者グループを発掘できる可能性を意味している．しかし，今回対象とした研究者 A，B に対しては，二人を媒介する第三の研究者が本大学内に存在したが，一般的にはこのような研究者が一つの組織内で求められる可能性は低い．

2.2 第三の研究論文を用いた方法

そこで，2 つ目の抽出方法として，媒介となる第三の研究者の代わりに，二人の研究者の異なるトピックを結びつける研究論文（根拠論文）が存在するのであれば共同研究の可能性があると仮定した方法を考案した．

この場合，根拠論文は前もって収集した組織内の研究者の論文ではない．しかし，すべての論文を収集して Author-topic-model で分析するのは非現実的であり，論文検索データベースを利用することになる．このため，Author-topic-model で分析して得られたトピックを含む論文の検索をキーワード検索で実現する必要がある．トピック t に対して，t から語 w が生成される確率 $\phi_w^{(t)}$ が高い語 w を検索キーとして論文を検索することが考えられるが，これには 2 つの問題がある．

一つは，$\phi_w^{(t)}$ が高い語 w がトピック t に特徴的な語とは限らず，w は様々なトピックで出現する場合があり，$\phi_w^{(t)}$ が高い語 w を検索キーとして検索した結果得られる論文が，トピック t を含んでいるとは限らない．もう一つは，語 w がトピック t に特徴的な語である場合でも，1 語だけを検索キーとして検索した場合は，検索結果の論文の中には，たまたま語 w を含むだけでトピック t とは関連しない論文も含まれる可能性が高いことである．

前者の問題を解決するために，以下のエントロピーを導入し，

$$H(w) = -\sum_t P(t|w) log P(t|w)$$

エントロピー $H(w)$ が閾値以下の語で $\phi_w^{(t)}$ が高い語 w を検索に用いる．語 w が生成されている場合に w がトピック t から生成された確率 $p(t|w)$ はトピック分析の結果得られる，各語に割り当てられたトピックを利用して推定する．また，後者の問題を解決するため，1 語ではなく，複数の語からなるクエリを用いて検索する．具体的には，$H(w)$ が 3 以下の語 w のうち $\phi_w^{(t)}$ の値の上位 10 個の語から複数の語を組み合わせた AND クエリを生成し，これによりトピック t を含む論文の検索を行う．今回用いたデータで「流体の可視化」に関すると思われるトピックを構成する語の生成確率とエントロピーを表 1，2 に示す． 表 1 のエントロピーに制限を設けない場合では上位 10 件に「計測」「速度」「濃度」「相関」「ベクトル」といった他の分野にも頻出すると考えられる語が出現確率の上位にきている．一方エントロピーに制限を設けた表 2 の場合では「LDA」「可視化」「シーディング」「マトリックス」「気流」のようにトピックに関係が深いと考えられる語が上位 10 件に入り込んでいる．

表 1: エントロピーに制限を設けない場合

W	P	H
計測	0.030141	3.673118
乱流	0.026884	2.299498
流れ	0.025573	3.734699
速度	0.017111	3.793496
羽根	0.015726	1.089181
濃度	0.015314	3.932315
相関	0.014416	3.687638
画像	0.014079	1.678916
LDA	0.012918	0.306400
ベクトル	0.012431	3.265508
測定	0.012169	4.114747

表 2: エントロピー 3 以下の場合

W	P	H
乱流	0.026884	2.299498
羽根	0.015726	1.089181
画像	0.014079	1.678916
LDA	0.012918	0.306400
可視化	0.011570	2.896868
シーディング	0.010896	0.123278
マトリックス	0.008762	2.762026
輝度	0.006553	2.282500
気流	0.006066	2.108830
HW	0.005766	1.084220

上記のようにして生成されるトピック t を含む論文の検索のためのクエリの集合を $Q(t)$ で表し，クエリ q を用いた検索結果の論文集合を $R(q)$ で表すと，研究者の組 (A，B) が以下を満たす場合，共同研究の可能性があるものと判断する．

$$\exists j\ \exists k\ \exists q_1 \in Q(j)\ \exists q_2 \in Q(k) \left[\theta_j^{(\mathrm{A})} \geq T\ \&\ \theta_k^{(\mathrm{B})} \geq T\ \&\ R(q_1) \cap R(q_2) \neq \emptyset\right]$$

結果は，(A，B) だけでなく，上記のトピック j，k，用いた検索クエリ q_1, q_2，および根拠論文として $R(q_1) \cap R(q_2)$ も出力する．最終的には，出力される根拠論文が 2 つの検索クエリによる検索結果として妥当か否かを判断し，妥当であれば (A，B) は共同研究の可能性があるということになる．

3 実験および結果

九州大学学術情報リポジトリ（QIR）から日本語で記述された pdf データを取得し，pdf2text を利用して

テキストに変換した．QIR には論文以外の著作物も存在するため，予備的に学術論文のテキストサイズを調査し，その結果からテキストファイルのサイズが 8 キロバイト〜45 キロバイトのものを学術論文と想定して取り出した．取り出した論文は 10422 件である．QIR の論文と論文著者の id が一対一に対応付けられたテキストデータ情報を利用して論文と著者の関係を抽出した．論文テキストファイルは日本語 Wikipedia エントリ を辞書に登録した MeCab を用いて形態素解析し，Wikipedia エントリのみを残し，他の語は削除して，論文データを作成した．Wikipedia エントリに限定した理由は２つある．１つはストップワードを除去するためである．獲得したデータに含まれる全ての語を実験データとして用いた場合，分かち書きされた語の中には明らかに意味を成さないような特殊記号や指示代名詞などが多数出現したからである．２つ目の理由は Wikipedia に登録されていない単語が著者を特徴付けるようなキーワードになっている可能性は低いと考えたからである．Author-topic-model に基づいて論文データをトピック分析し（トピック数 600），著者（研究者）ごとのトピックの出現確率分布，トピックごとの語の出現確率分布を得た．

まず，論文データ中の論文著者で，九州大学に所属する研究者をすべて抽出した．抽出した全研究者から分野は問わずランダムに 30 名抽出し，これから研究者の組 435 組を生成し，分析対象とした（分析対象 1）．また，共同研究の可能性が高いと考えられる研究分野である「認知科学分野」「情報系分野」「脳神経科学分野」に限定して研究者をランダムに 10 名取り出し，研究者の組 45 組を生成し分析対象とした (分析対象 2)．

今回は，著者に有意に表れるトピックを求めるための閾値 T を $T = 0.2$ と設定して実験を行った．また，トピックを構成する語から生成するクエリは，2 語で構成するようにした（3〜4 語で構成したクエリでも試してみたが，このようなクエリによる検索は失敗したため）．

論文検索データベースとしては CiNii を利用し，API を用いて検索した．分析対象 1 の研究者の組 435 のうち，145 組が共同研究の可能性があるものとして抽出された．抽出された研究者の組に対し，九州大学研究者情報データベースに記載されている両研究者の研究活動の状況と，同時に出力される根拠論文に基づいて判断したところ，4 組が共同研究の可能性が高いと思われた．その一例を表 3 に挙げる．

分析対象 2 の研究者の組 45 のうち，3 組が共同研究の可能性があるものとして抽出された．分析対象 1 の場合と同様に，研究活動状況および出力される根拠論文に基づいて判断したところ，2 組が共同研究の可能性が高いと思われた．その一例を表 4 に挙げる．

多くの場合は，可能性ありとして出力される組に対して，その研究者の研究活動の状況を調べれば共同研究の可能性があるか否かを判断できる．一方，分析対象 2 の結果として挙げた表 4 の例では，認証技術と画像処理は一見関連しないように思われ，研究活動状況を見ても共同研究が成立すると判断できる人はそう多くはないであろう．しかし，根拠論文として挙げている論文の書誌情報（タイトル，アブストラクト）を見ることで，どうして共同研究につながるのかが理解できる．根拠論文の書誌情報は，単に，用いたクエリーの検索結果として出力された論文が妥当なものか否かを判断するために利用するだけでなく，共同研究が成立する理由をシステム利用者が理解するのにも利用できる．

表 3: 発掘された著者対と共同研究の可能性を示す論文タイトル

共同研究の可能性を示唆する論文タイトル	
[1] 持久的競技者におけるヘモグロビン遺伝子変異と酸素結合能に関する研究 (酸素 健康/遺伝子 変異) [2] 疾病予防・健康増進のための分子スポーツ医学 (13) 肥満と減量に対する遺伝的要因 (肥満 健康/遺伝子 塩基) [3] β 3 アドレナリン受容体遺伝子解析を応用した減量指導 (肥満 健康/遺伝子 変異 DNA) [4] 健康管理における個人素因の重要性と体質診断としての一塩基遺伝子多型の可能性 (肥満 健康/遺伝子 変異 塩基)	
研究者 1	活動
	実験室での研究として，1）長時間運動後の代謝高進のメカニズムを明らかにするため生理的および生化的的な面から検討している．また 2）運動強度および時間の抱合型カテコラミンに及ぼす影響を検討している．フィールド研究として，ネパール等で現地調査を実施し，生活形態と健康態に関して広領域的国際比較疫学研究を実施している．文部科学省 21 世紀 COE 拠点形成プログラム「循環型住空間システムの構築」では，建設作業における作業特性と生体反応に関する研究として，作業時の生体負担度や作業姿勢とバランスに関して研究した．キーワード：長時間運動， 代謝基質， ホルモン， 呼吸循環， ネパール， 生活習慣病疫学， 身体活動量， 健康度， 抱合型カテコラミン， 季節変動
Topic 語	酸素 肥満 健康 心拍数 脂肪
研究者 2	活動
	抗ウイルス・抗ストレス・抗アレルギー・免疫増強・抗酸化・抗生活習慣病効果をもつ機能性食品の開発，プロモーター活性化による動物細胞の組換えタンパク質高生産性細胞株の樹立法の開発，無血 清・無タンパク高密度培養法の開発，体外免役法による抗原特異的ヒトモノクローナル抗体の作成とその機能改変を行っている．
Topic 語	DNA 遺伝子 変異 蛋白質 染色体 染色 塩基 プラスミド 塩基配列 残基

タイトルの（ ）内の語は（著者１が使用したクエリ/著者２が使用したクエリ）を表す

4 考察

今回の実験で分かった問題点は検索クエリを決定する生成確率とエントロピーの設定が一意に決まらない，

表 4: 発掘された著者対と共同研究の可能性を示す論文タイトル

共同研究の可能性を示唆する論文タイトル	
[1] 動画像によるオンライン署名認証: Sequential Monte Carlo を用いたペン先追跡 (システム 認証 Web/画像 カメラ システム)	
研究者 1	活動
	専門分野情報科学，分散システム，情報検索，Web サービス，電子認証・認可活動概要●研究- コンテンツ検索，情報検索，Web マイニング- クラウドコンピューティング- 図書館の電子サービス・機関リポジトリ- Web 情報サービス構築●教育- 大学院システム情報科学府での教育を担当．- 大学院の講義および修士研究を担当．●職務- 情報基盤研究開発センターの教員として，情報統括本部が行う学内情報サービス基盤の構築および運営を行う．- 学内のクラウド計算環境，全学共通認証基盤，IC 職員証・学生証，全学基本メールを管理．- 九州大学附属図書館研究開発室の研究員として，図書館サービスの研究開発．
Topic 語	システム 認証 サーバー コミュニティ 情報サービス Web
研究者 2	活動
	専門分野コンピュータビジョン活動概要【研究業績】画像処理，コンピュータビジョン，並列処理システム，マルチメディア情報処理の研究に従事し，それらの研究成果を論文等に執筆する一方，画像処理ソフトウェアシステムの開発に従事．これらの研究成果により，電子情報通信学会篠原記念奨励賞 (1989)，情報処理学会論文賞 (1993)，情報処理学会坂井記念特別賞 (1995)，映像情報メディア学会丹羽高柳論文賞 (2001) を受賞．電子情報通信学会フェロー (2009)．情報処理学会フェロー (2013)．また，情報処理学会，電子情報通信学会，電気学会，映像情報メディア学会，画像電子学会，IEEE などにおいて，各種委員を歴任した．【教育活動】大学院システム情報科学府情報知能工学専攻，工学部電気情報工学科の授業を担当．また，非常勤講師として，熊本大学大学院，福岡大学大学院の授業を担当した．【社会連携活動】企業，地方自治体からの受託研究や共同研究を実施する一方，研究アドバイザーやコンサルティングを行う．詳細は別途項目を参照のこと．
Topic 語	画像 カメラ システム 画素

タイトルの ( ) 内の語は (著者 1 が使用したクエリ/著者 2 が使用したクエリ) を表す

ないしキーワードの選択が Wikipedia の見出し語だけでは対処できない場合が生じる点である（本実験では Wikipedia の見出し語をドキュメントの素性として用いた）．例えば「画像処理技術を応用した手術支援」に関する論文が根拠論文として検索されるのが望ましい場合,「画像処理技術」「手術支援」といったキーワードが望ましいが，条件によってはどちらか一方のキーワードが条件に当てはまらない場合がある．また,「クエリ拡張」といった複合名詞が有効な検索クエリになる場合にキーワードが「クエリ」「拡張」と別 id で登録されていればこれらも条件に当てはまらない場合がある．つまり特定の語同士の組み合わせになると有効なクエリとなり得るが，それを構成するキーワードだけに注目した場合はありふれたキーワードとして解釈され有効なクエリが構成できない場合がある．

今回のシステムでは共著論文候補 145 組に対して実際に共同研究の可能性が考えられる論文数が 4 組と精度が低い結果となった．これは設定したトピック数 (600) が実際のデータに対しては少なく，著者の興味を十分に表現できる粒度ではなかったことが考えられる．しかし，今回の実験で用いたトピック数でも収束するまでに数週間を要したため，システムの高速化も今後の課題である．

5 まとめ

学際的な研究の促進において重要と考えられるリサーチ・アドミニストレータの業務遂行補助を目的として，収集した組織内の研究者の論文集合を Author-topic-model で分析し，この分析結果と一般的な論文検索データベースを利用して，共同研究の可能性がある研究者の組の候補を抽出する手法を提案した．また，この手法を用いて小規模な評価実験を行った．

可能性ありとして抽出される研究者の組のうち，共同研究の可能性が高いものはまだ非常に少なく，考察でも述べたような改良が必要である．一方，事前に共同研究が発生しやすいと考えられる分野に限定した実験では，根拠として提示された論文を確認することではじめて共同研究の可能性を確認できる実験結果が存在した．この結果は提示される論文データが共同研究の可能性を説明する一助になることを示していると考えられる．

今回の実験では目視による主観的な評価にとどまったが今後実際にリサーチ・アドミニストレータとして研究の企画立案に携わっている職員に研究結果を評価してもらう予定である．

謝辞

本研究の一部は JSPS 科研費 25540151 の助成を受けて行ったものです．

参考文献

[1] Michal Rosen-Zvi and Thomas Griffiths and Mark Steyvers and Padhraic Smyth. *The Author-Topic Model for Authors and Documents*. UAI '04 Proceedings of the 20th conference on Uncertainty in artificial intelligence Pages 487-494

[2] Michal Rosen-Zvi and Thomas Griffiths and Mark Steyvers and Padhraic Smyth. *Learning Author Topic Models from Text Corpora*. ACM Transactions on Information Systems (TOIS) Volume 28 Issue 1, January 2010 Article No. 4