公開

（事後評価・研究者用）

研　究　報　告　書

「学術データの自動集約による研究者プロファイリング」

研究期間： 2018 年 10 月～2020 年 3 月
研究者番号： 50172
研 究 者： 桂井　麻里衣

1.　研究のねらい

個々の研究者が生産した成果（例：学術論文，グラント，受賞，特許，学位論文）を，その種別を問わずに集約する研究者プロファイリングは，個人や機関の評価，学術コミュニティ全体の評価といった「科学の科学（science of science）」を推進するうえで重要な基礎技術となる．現在，国内外には様々な種類の学術データベースが存在しており，各々が独自に著者の個人識別子（以降，著者 ID）を用意して学術情報を管理している．データベースを横断した著者 ID は未だに整備されておらず，一人の研究者に関する多様な成果は複数データベースに散在した状態である．ゆえに，単純に研究者氏名を検索キーワードとして業績を収集すると，同姓同名研究者の存在が問題となる．本研究のねらいは，複数の学術データベースから個人の情報を正確に収集するための手段を提供し，学術データベース分析に関する研究開発や実務に役立てることにある．

上記の研究開発により，様々な分野の研究者のプロファイルが完成する．その情報検索への一応用として，プロファイルから推定した専門内容に基づく関連研究者推薦モデルを開発する．異分野の知識を融合させた研究は，既存の学問分野の枠組みをこえた発想や技術を生み出すといわれており，国内でも学際研究推進のために様々な政策的努力がなされている．しかし，これまでに提案者らの研究では，国内の大学・研究機関において異なる部局の研究者らによる共同研究が非常に少ないことを示した．これは分野外の研究者を知る機会の少なさが問題といわれている．そのため，研究者プロファイル間の類似性を発見する手法を構築し，異分野コミュニケーション機会を誘発するような新たな検索フレームワークを提供する．

2.　研究成果

（1）概要

学術データの著者同定については従来盛んに研究されてきたが，いずれも学術論文のような単一のデータを対象に設計されている．その多くは共著関係や引用関係のような学術論文固有の特徴に依存しており，論文以外の学術データから著者同一性を発見するタスクに適用することが困難である．その他，テキスト内容の類似性検証に基づく著者同定手法も提案されているが，日本語や英語など異なる言語で書かれた論文の著者をリンクする手法は未だ存在しない．これらの問題を解決するために，本研究では，(i) 学術データの種類を問わずに利用可能な特徴に基づく著者同定手法と，(ii) 異なる言語間で著者同一性を発見する技術をそれぞれ開発した．さらに，提案した集約技術の応用として，(iii) 関連研究者の検索インタフェースを開発した．

（2）詳細

概要で述べた研究項目 (i)〜(iii) について，それぞれ詳細を説明する．

(i) 異なる種類の学術データの著者マッチング

種類の異なる学術データベースが二つ与えられたとき，データベースを横断して著者 ID を紐付ける問題に取り組んだ．本研究では，データベース管理者による手動マッチングの労力低減を目的とした，人間参加型フレームワークを提案した．具体的には，一方のデータベースの著者 ID がクエリとして与えられたとき，他方のデータベースにおける同姓同名著者を同一人物らしさでランキングする．はじめに，著者氏名文字列の比較に基づき，一方のデータベース内著者に対する他方の同一人物候補集合を構築する．次に，多くの学術情報データベースで入手可能なメタデータ(タイトル，所属，共著者情報)のみが利用できるという制限のもとで，複数の類似度尺度を算出する．それらの教師なし集約によって算出した最終スコアに基づき，同一人物らしいと考えられる順に同姓同名著者をソートする．

提案した類似度尺度は本研究の新規性の一つである．著者同定の従来研究では，共著者氏名の重複度は強力な特徴とみなされてきた．しかし，博士論文のような単著の文献については，共著者氏名の重複を算出することが不可能である．そこで，一方の学術情報に共同研究者が存在する場合，他方の学術情報の著者との所属類似度を算出した．これにより，著者本人の所属が変わっていたとしても，以前の研究機関に共同研究者を残している場合を考慮できる．その他，所属文字列に基づく類似度，発表年に基づく類似度，研究内容に基づく類似度を提案した．

提案手法の有効性を評価するため，CiNii Dissertations 内の博士論文著者と，KAKEN 研究者 ID のマッチング実験を行った．同姓同名者数が上位 8 件となる氏名に対する結果と，手動で用意した正解データと比較することで，マッチング正解率を算出した．類似度尺度を単体で用いた場合に比べ，教師なし集約することで正解率が向上することを示した．これらの成果は国際会議 JCDL2019 で発表した（研究業績[1]）．以上のように，著者マッチングを全自動化するのではなく，人間参加型のデータベース運営技術として提案することで，研究目的はおおむね達成できたといえる．現在は実験内容を追加し，ジャーナル投稿論文を準備中である．

(ii) 異なる言語の学術データの著者マッチング

(i) を複数言語へと拡張した問題である．言語の異なるデータベース間で著者を対応付けるには，氏名比較に基づく同一人物候補者の抽出と，メタデータや研究内容に基づく類似度算出の二点を言語横断的に実装する必要がある．そこで本研究項目では，漢字からローマ字氏名への変換，アルファベット表記を含めた同一人物候補者の抽出，翻訳に基づく研究テキスト間の類似度算出を新たに追加した．KAKEN の研究者 ID と DBLP 文献情報のマッチング実験により，単一の類似度を用いた場合の性能と複数の類似度を統合した場合の性能を比較したところ，共著者情報とテキスト内容を統合した類似度の性能が最も高い性能を示した．同時に，発表年に基づく類似度算出指標については検討の余地があることを確認した．

本研究は日本語と英語のマッチングを対象としたが，提案手法は文字列変換の部分を変更することで様々な言語に適用可能であり，研究目的はおおむね達成できたといえる．以上の成果は国内学会 DEIM2019 で発表した（研究業績[7]）．現在は実験内容を追加し，ジャーナル投稿論文を準備中である．

(iii) 異分野関連研究者推薦

構築した研究者プロファイルの応用として，様々な分野の関連研究者を推薦する手法を構築した．共同研究者推薦の従来研究では，ユーザが関連研究者を発見できるシステムを実際に構築した例は少ない．そこで本研究では，様々な分野の関連研究者を地図上で効率的に発見可能な検索・可視化システムを新たに構築した．KAKEN 研究者データセットを構築し，研究者間の関連度はキーワード・課題タイトルなどの単語埋め込みに基づくベクトル間類似度として算出した．

提案システムは，研究者や研究機関，キーワードをクエリとして，それらの関連研究者を図 1 に示すように日本地図にマッピングする．研究者をクリックすると，図 2 に示すように，専門内容や過去の科研費採択課題，氏名と所属，関連するウェブ画像を表示する．実データを用いたシミュレーションにより，システムの提示する関連研究者は，既存の共同研究関係から導かれる関連性との重複が少なく，新たな共同研究者候補の発見支援が可能となることを示した．以上の成果は国内学会 DEIM2020 で発表した（研究業績[4]）．今後はさらに機能を追加する予定である．

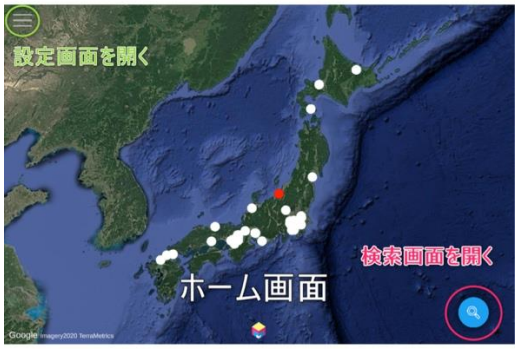

図 1 提案システムのホーム画面．

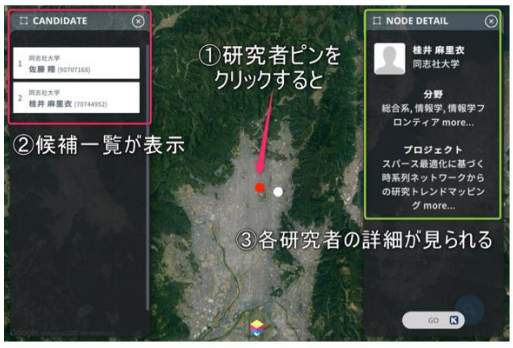

図 2 研究者の詳細画面．

3. 今後の展開

本研究は，学術ビッグデータ分析基盤の整備を目的としたものである．研究者の成果に関する構造化データを提供することで，データマイニングや機械学習技術の適用が容易となり，「科学の科学」の研究のさらなる加速が見込める．今後は，研究トレンドの自動抽出手法（研究業績[2, 5, 6]）と組み合わせることで，研究者の興味変化の分析へと発展させることを考えている．

研究者の業績を一元管理することの重要性を示すことで，複数データベース間の連携が進むと考えられる．世界中の研究者情報を集約・整備すると，海外の関連研究者推薦へと発展でき，国際コラボレーション創出につながる．加えて，本研究のように様々な分野の研究者の履歴を可視化することは，若い世代の進路選択にも役立つ．研究項目(iii)（研究業績[4]）で構築した関連研究者システムの実用化を進めることで，研究者に対してのみならず，市民と科学の架け橋を実現したい．

将来的には，適用対象を研究者にとどめず，本研究で固めた基礎技術を産業界など別のドメインへ応用展開する．ビジネス間や各種技術との潜在的なつながりを分析するための土台を提供することで，産学の大規模なジョブマッチングを実現し，新規ビジネスやコンセプトの開拓につなげていきたい．

4. 自己評価

1 年 6 ヶ月間において全ての研究項目に取り組み，その成果を国内外で発表したことから，当初の研究目的はおおむね順調に達成できたといえる．特に研究業績[1]の発表は，研究業績[3]の招待論文執筆へと発展できた．今後は実験を追加して手法の問題点を考察し，さらなる改良を重ねる予定である．

本研究は，国立情報学研究所の大向一輝准教授（現：東京大学准教授）から研究用データの提供を受けて実施した．同志社大学の学生を研究補助者に加え，計画通りに研究費を執行した．

異なる学術データベースの著者マッチングという課題は新しく，特に言語横断という挑戦性があった．関連研究者推薦の従来研究では，いずれもコンピュータサイエンスという特定分野にとどまっていたが，本研究はあらゆる分野を包括した大規模学術データベースから情報集約したため，分野横断型の検索を実現した．

研究者の業績集約は，「科学の科学」を推進するための基盤技術である．研究開発用リンクト・オープン・データの推進に加え，分野内に閉じた知識を可視化し，新たな研究開発のための意思決定支援につながる点が意義となる．また，異分野の研究者らの協働促進は，図書館情報学のみならず社会学の知見・方法論が必要不可欠であり，情報学という立場からこれらの分野にデータ駆動型アプローチを提供することも貢献の一つと考えている．

5. 主な研究成果リスト

(1)論文(原著論文)発表

1. Marie Katsurai and Ikki Ohmukai, “Author Matching Across Different Academic Databases: Aggregating Simple Feature-Based Rankings,” Proceedings of 2019 ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL2019), pp. 279-282, 2019. (査読あり)
2. Marie Katsurai and Shunsuke Ono, “TrendNets: Mapping Emerging Research Trends From Dynamic Co-Word Networks via Sparse Representation,” Scientometrics, vol. 121, pp. 1583-1598, 2019. (査読あり)

(2)特許出願

研究期間累積件数:0 件

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

3. Marie Katsurai, “Using Word Embeddings for Library and Information Science Research: A Short Survey,” ACM SIGWEB Newsletter, Winter Issue, 2020, to appear. (招待あり)
4. 丹後綱也, 西澤浩之, 近澤悠登, 桂井麻里衣,「関連研究者と所属位置情報の検索・可視化システム」, 第 12 回データ工学と情報マネジメントに関するフォーラム(DEIM2020), A7-4, 2020 年 3 月.
5. 桂井麻里衣, 小野峻佑,「語の共起関係に基づく研究トレンドの可視化」, 電子情報通信学会技術研究報告, vol. 119, no. 185, pp. 23-27, 2019.
6. 電子情報通信学会基礎境界ソサイエティ信号処理研究会「令和元年度 信号処理研究会賞」受賞(受賞対象:研究業績[5]), 2020 年 3 月.
7. 近澤悠登, 桂井麻里衣, 大向一輝,「言語の異なる学術データベース間の著者同定」, 第 11 回データ工学と情報マネジメントに関するフォーラム (DEIM2019) , I4-2, 2019 年 3 月.