E-031

# researchmap の情報を用いた異なる研究分野間の研究者マッチングの試み
# Attempting Researcher Matching across Different Research Fields Using Information from researchmap

前多 陸玖[1)]　廣吉 和貴[1)]　木村 泰知[1)]
Riku Maeda　Kazuki Hiroyoshi　Yasutomo Kimura

## 1　はじめに

近年，学術論文の数が増加し，論文に関する情報（研究分野，研究課題，研究者など）を俯瞰してみるためにも，コンピュータによる解析が必要になっている．国内外の論文データベースの整備が進んでおり，コンピュータによる解析もさかんに行われている [1]．例えば，arXiv[1)]では，コンピュータサイエンスをはじめ，物理学，数学，統計学などの論文を閲覧できる．また，PubMed[2)]は生命科学および生物医学の書誌データベースの MEDLINE を検索できるサイトである．日本においても，CiNii Research は，日本の論文などの研究データを学術情報検索サービスでとして提供してる[3)][2]．さらに，これらの論文データベースは Web API が提供されており，プログラムを用いて収集できる．自然言語処理分野では，このような論文データベースを用いて，論文固有の文書構造を考慮した文書類似度，引用文脈の解析，論文からの情報抽出，論文の要約，論文についての質問応答などの論文解析タスクに取り組んでいる [3].

論文は解析するだけにとどまらず，学術論文データベースに記載されているデータベースから，共同研究者の情報などを用いて研究者コミュニティを発掘することにも利用できる [4, 5]．研究者の論文データベースは，異なる研究領域における新たな研究チームを作成するという視点でも重要な情報である [6]．近年は，異なる分野の研究者を結びつけ，異分野連携，あるいは，文理融合などの新たな融合領域を発見することが期待されている．しかしながら，専門外の領域との接点を見つける作業には，研究者の他学問に対する認知力に大きく依存している [3]．また，デジタル化による情報社会の中では，研究者の認知の限界を解消する手段が必要である [7].

そこで，本研究では，異なる分野の研究者の連携を促進することを目指し，researchmap に掲載されている論文タイトルを用いて，研究者の興味関心のある分野を推定し，興味関心の近い研究者を提示することを目的とする．具体的には，北海道国立大学機構として統合された商学の小樽商科大学，農学の帯広畜産大学，工学の北見工業大学に所属する研究者を対象として，各研究者が発表した論文のタイトルで類似度をとり，類似度が高い論文を書いている研究者のマッチングを試みる．

本研究の貢献は，下記の 3 つである．

- 異なる研究分野（商学，農学，工学）の研究者が所属する 3 つの大学を対象として，論文の類似度をにより，研究者マッチングを試みた．
- researchmap を扱う際の問題として，登録情報，あるいは，所属変更のタイムラグによる問題などについて明らかにした．
- 研究者マッチングの正解として，論文タイトルから同じ分野の研究者をみつけられるのか，あるいは，科研費の共同研究者になっていることを評価軸として検討した．

1）　国立大学法人 北海道国立大学機構 小樽商科大学
1）　https://arxiv.org/
2）　https://pubmed.ncbi.nlm.nih.gov/
3）　https://cir.nii.ac.jp/

## 2　関連研究

論文タイトルを用いて文章間の類似度を測るため関連研究には，論文のクラスタリングを [8] や類似論文検索 [9] などがある．

榊らの研究では，論文をトピックごとにクラスタリングする手法として，4 つの手法（共著関係，構造同値，著者論文，著者特徴語）で論文のネットワークを構築し，クラスタリングを行い，その評価を行った [8].

小林らの研究では，「目的は異なるが手法は似た論文」のような論文の基本的な構成要素である目的・手法・結果に基づいた論文検索の実現を目的に，論文の列挙共引用関係に注目した列挙共引用予測というタスクを提案した [9].

## 3　研究者マッチングとは

本研究における研究者マッチングとは，異なる研究分野 (商学，農学，工学など) から，興味関心の近い研究者をみつけることである．

ここで，研究者マッチングの定義について説明する．研究者のマッチングが成功する例としては，複数の視点から考えることが可能である．例えば以下のような観点が挙げられる．

1. 既存の研究成果に結び付けられるのか
2. 同じ分野であることを正しく評価できるのか
3. 研究以外のコミュニティ（SNS）で近いところに興味を持っている

研究者のマッチングをするためには，twitter で共通の人をフォローしていたり，異なる分野の研究者と研究補助金を獲得している，研究分野が同様であるなどの観点を用いることで，より実用的なマッチングを評価することができると考えられる．

## 4　実験設定

### 4.1　利用できるデータとは

本実験では，researchmap に掲載されている研究者情報を用いる．researchmap とは研究者が業績を管理・発信できるようにすることを目的とした，データベース型

研究者総覧である[4]．

本実験では，2022 年 10 月 25 日時点の researchmap から researchmap.v2 API（以下 rmap.v2 API）を利用して取得した小樽商科大学，帯広畜産大学，北見工業大学に所属する研究者の研究者情報を対象として行う．対象の研究者データ取得に際する課題として，退職者，転職者，新規採用者，を考慮しなければならない点が挙げられる．本研究は研究者のマッチングを目的としているため，すでに対象としている大学に所属していない場合を考慮しなければならない．この問題に対応するため，各大学の研究者総覧の情報も活用した．

データセット作成の具体的な手順は，次のとおりである．まず，researchmap で各大学の機関コードで検索し，表示された Web ページから Web スクレイピングの手法を用いて，各研究者の氏名とページ URL 末尾に設定されているパーマリンクを取得する．rmap.v2 API から取得できるデータは json 形式となっており，ここから Web スクレイピングを用いて必要とされるデータを選別した．次に，各大学の研究者総覧の Web ページから同様にして取得した研究者氏名と researchmap から取得した研究者氏名を照合し，研究者総覧に存在する研究者の氏名とパーマリンクのみを csv 形式で出力する．照合結果は，表 1，表 2，表 3 に示す．さらに，作成した csv ファイルから取り出したパーマリンクを rmap.v2 API の URL と結合し，その URL にアクセスすることで取得した研究者情報を JSON 形式で出力する．

マッチングに必要なデータとしては paper_title（論文・misc のタイトル），total_items（論文・misc の総数），publication_date（発行日）などがある．researchmap 上の研究者は，論文よりも misc の掲載数が多い傾向にある．また，論文と misc の区別は個人の裁量で決められており，国際会議などは研究者ごとに判断が異なっていた．そのため，マッチング対象の論文として，misc も含めることで，より有効なマッチング成果が期待できる．また，paper_title が英語の場合は，DeepL を使用し日本語に翻訳した．例えば，「Arthrectomy for traumatic proximal interphalangeal arthritis in the lateral digit in a heifer」という英語タイトルの場合は，「未経産牛の外傷性近位指節間関節炎に対する関節切除術」と翻訳され，類似度が計算される．また上記の例から，専門用語に関しても問題なく翻訳可能であることがわかる．タイトルが英語であるかどうかの判定は，タイトルの文字が ASCII 範囲の 127 以内であるかどうかで判断した．そのため論文タイトルのすべてがアルファベットで構成されいる場合は翻訳され，タイトル内に日本語とアルファベットが混在している場合には翻訳されない．翻訳機を使用することでの英語論文と著者マッチングの性能の安定性は，近澤氏の研究 [10] で確認されている．

本実験で対象とした各大学の研究者の数を表 1，表 2，表 3 に示す． researchmap から取得した 3 大学（小樽，帯広，北見）の論文，および，misc の数を表 4，表 5，表 6 に示す．図 1 に 3 大学の年度別の論文タイトル数と評価実験で対象とするタイトルを示す．論文タイトルの収集期間は 1983 年から 2024 年までである．本実験では，研究者マッチングに活用するため，2018 年から 2024 年までの論文タイトルを対象とする．

**表 1　小樽商科大学の照合結果**

| 小樽 | rmap 上に無し | rmap 上に有り |
|---|---|---|
| 総覧に氏名無し | - | 20 |
| 総覧に氏名有り | 14 | 104 |

**表 2　帯広畜産大学の照合結果**

| 帯広 | rmap 上に無し | rmap 上に有り |
|---|---|---|
| 総覧に氏名無し | - | 59 |
| 総覧に氏名有り | 30 | 105 |

**表 3　北見工業大学の照合結果**

| 北見 | rmap 上に無し | rmap 上に有り |
|---|---|---|
| 総覧に氏名無し | - | 78 |
| 総覧に氏名有り | 33 | 96 |

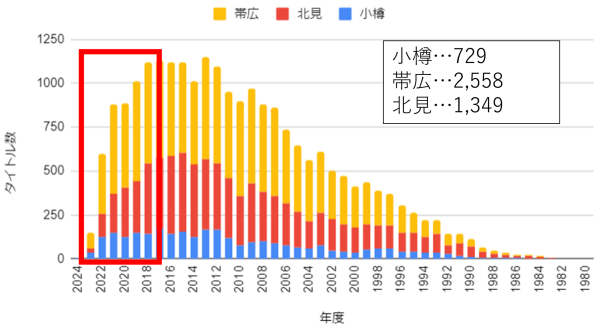


**図 1　3 大学の年度別の論文タイトル数.**

## 4.2　研究者マッチングの手法

下記に，研究者マッチングの流れの示す．

**Step1** researchmap の「論文タイトル」と「共著情報」を収集する収集後，英語論文誌，国際会議などの英語タイトルのみの場合には，DeepL の API を用いて，英語から日本語へ翻訳する．

**Step2** 研究者ごとに論文タイトルのリストを作成する

**Step3** ある研究者 A の論文タイトルリストを入力として，他の研究者の論文タイトルをペアとして類似度の結果を付与したリストを出力する．この際に，共著になっている論文タイトルは対象外とする．例えば，八久保氏, 舘山氏は，下記の 2 つのタイトル論文を共著で書いている．

1.「塩を含む積雪の含水率測定に関する考察」
2.「塩を含む積雪の含水率測定法」

上記の 2 つの論文は，両者の論文タイトルリストにそれぞれ含まれている．このような共著論文が存在した場合には，両者のペアのリストを出力結果から除外する．

**Step4** ある研究者 A の論文タイトルリストに対して出力されたリストから，研究者ごとに研究者 A と類似度の最も高い論文タイトルの類似度スコアを代表スコアとしたリストを作成する．

文書間類似度は TFIDF COS 類似度を用いて算出した．研究者ごとに作成した論文タイトルリストを，すべて統

4）https://researchmap.jp/

表 4　小樽商科大学の論文・misc

| 小樽 | 平均値 | 最大値 | 最小値 | 中央値 | 合計 |
|---|---|---|---|---|---|
| 論文 | 15.8 | 103 | 0 | 11 | 1,641 |
| misc | 12.5 | 125 | 0 | 4.5 | 1,305 |
| Total | 28.3 | 172 | 0 | 19 | 2,946 |

表 5　帯広畜産大学の論文・misc

| 帯広 | 平均値 | 最大値 | 最小値 | 中央値 | 合計 |
|---|---|---|---|---|---|
| 論文 | 63.1 | 420 | 0 | 43 | 6,627 |
| misc | 50.4 | 378 | 0 | 27 | 5,295 |
| Total | 113.5 | 696 | 0 | 74 | 11,922 |

合した．その後統合した論文タイトルリストの特定のタイトルを key として，その他の論文との類似度を求める．TFIDF を行う前の処理として，論文タイトルから非正規表現を用いて html タグを取り除く作業を行った．

1.「CH<sub>3</sub>D と CD<sub>4</sub> を包接した同位体メタンハイドレートの相平衡について」
2.「CH3D と CD4 を包接した同位体メタンハイドレートの相平衡について」

上記のような作業により，html タグを含んだ論文同士の類似度が高くなるのを防いだ．TFIDF を行う上で，各論文タイトルを Janome を用いて形態素解析を行い，名詞のみを抽出した．その後 TFIDF を scikit-learn を用いてベクトル化し，論文同士のコサイン類似度を算出した．またコサイン類似度を算出する際は，研究者自身の論文同士の類似度を求めないようにした．文書間類似度を算出したのちに，研究者同士の代表スコアを算出した．

## 5　評価実験

### 5.1　目的

本研究では，異なる研究分野，あるいは，異なる大学の間で研究者のマッチングがどの程度できるのか，明らかにする．

### 5.2　評価方法

本研究において，研究者マッチングを以下の方法で評価する．

- **評価 1** : 科研費の研究チームになっている研究者をみつけられた場合には正解とする．
- **評価 2** : researchmap の研究分野が同じ研究者をみつけられた場合には正解とする．

researchmap の研究分野は，研究者が自身の研究領域に基づいて設定することができる．rmap.v2 API を用いて，対象である研究者の研究分野を収集し，同じ研究分野である研究者をペアとする正解データセットを作成した．その正解データセットをから，研究者マッチングの再現率と適合率を算出し評価する．

### 5.3　結果

科研費による評価 (**評価 1**) に関しては，採択された科研費のプロジェクトにかかわる共同研究者を正解とみなすことを試みたものの，３大学内で共同研究者になっている科研チームは３件しかみつからなかったため，評価を行うことができなかった．そのため，本研究の評価は「researchmap の研究分野が同じ人をみつけられた場合 (**評価 2**)」を中心に行う．

表 6　北見工業大学の論文・misc

| 北見 | 平均値 | 最大値 | 最小値 | 中央値 | 合計 |
|---|---|---|---|---|---|
| 論文 | 45.7 | 223 | 0 | 30 | 4,389 |
| misc | 35.0 | 313 | 0 | 16 | 3,356 |
| Total | 80.7 | 503 | 0 | 46.5 | 7,745 |

#### 5.3.1　再現率

再現率はある閾値に対して，正解とするデータセットの研究者ペア数のうち，出力した研究者ペア数に正解のペア数がいくつ存在するかで計算する．類似度を閾値として設定し，0.0 から 0.9 までの推移を 0.1 刻みで計測した．

#### 5.3.2　適合率

適合率はある閾値に対して，出力した研究者ペア数のうち，正解とするデータセットの研究者ペア数がいくつ存在するかで計算する．類似度を閾値として設定し，0.0 から 0.9 までの推移を 0.1 刻みで計測した．

#### 5.3.3　F 値

算出した再現率と適合率をもとに F 値を求める．図 2 は適合率の推移である．

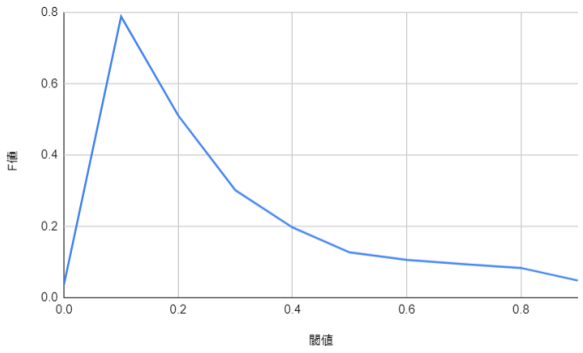


図 2　F 値の推移．

### 5.4　考察

実験の結果，閾値が 0.1 の場合の F 値が最も大きいことが分かった．その理由として，閾値が高くなるにつれ正解率が大幅に低下したことが挙げられる．また閾値が 0.8 以上の場合は，マッチング数が少なくなるため適合率の変動が大きいことが分かった．

本研究では，論文タイトルの類似度を求めることで，研究者同士のマッチングを行った．三大学を対象とした中で，他大学の研究者同士のマッチングが見られた．図 3 に示すように小樽商科大学の多田氏と帯広畜産大学の河野氏のマッチングを行った．どちらの論文も「消費者の購買行動」について書かれており，両者は同様の研究領域を有していると考えることができる．一方で，今回は TFIDF を主な手法として採用したため，同義語や類似語の類似度が低くなってしまった．そのため今後は word2vec や SentenceBERT などの使用を考える必要がある．

本研究では事前に rmap.v2 API を使用し，各論文ごとに共著者を収集し，共著者同士でのマッチングを行わないようにした．しかし，researchmap 上に存在する共著者名がローマ字表記である場合は，判別することができなかった．

例えば図 4 に示すように，古林氏の「ホルスタイン子牛に発生した単一左心室の一例」の論文を使用し，同論文の共著者である渡辺氏とマッチングした．これは researchmap に存在する共著者情報が，「Kenichi Watanabe」とローマ字表記であったことから判別ができなかったと思われる．

また本研究では DeepL を使用した．英語タイトルを翻訳したことでより幅広いマッチングを行うことができた．西口氏の「SuzuPoli: Political Party Matching App」を「すずぽり：政党マッチングアプリ」と翻訳したことで，図 5 に示すように西口氏と木村氏のマッチングを行うことに成功した．

| 研究者名 | 論文タイトル | 研究者名 | 論文タイトル | 類似度 |
|---|---|---|---|---|
| 多田伶 | 消費者の思考モードと購買意思決定: 消費者行動論における無意識的思考研究の応用 | 河野洋一 | エシカル消費に対する消費者の認知状況と購買行動の特徴 | 0.457 |

図 3　他大学の研究者マッチングの成功例.

| 研究者名 | 論文タイトル | 研究者名 | 論文タイトル | 類似度 |
|---|---|---|---|---|
| 古林与志安 | ホルスタイン子牛に発生した単一左心室の一例 | 渡邉謙一 | ホルスタイン子牛の単一左心室の臨床例。 | 0.792 |

図 4　共著者判別の失敗例.

| 研究者名 | 論文タイトル | 研究者名 | 論文タイトル | 類似度 |
|---|---|---|---|---|
| 西口純代 | すずぽり：政党マッチングアプリ | 木村泰知 | 東京都議会議事録における政党を用いたスタンス分類について | 0.251 |

図 5　翻訳による研究者マッチングの成功例.

## 6　まとめ

本研究では，異なる分野の研究者の連携を促進する方法として，researchmap に掲載されている論文タイトルを用いて，研究者の興味関心のある分野を推定し，興味関心の近い研究者を提示した．

異なる研究分野（商学，農学，工学）の研究者が所属する 3 つの大学を対象として実験した結果，類似度が 0.1 の場合の F 値が最も高いという結果となり，タイトルからの研究者マッチングの可能性を示した．

研究者データベースである researchmap を扱う際の問題点を述べる．researchmap を扱う際には，研究者の所属変更が頻繁に行われることから，所属変更のタイムラグを考慮する必要があった．また，異なる分野によって対象とする言語が異なるため，機械翻訳により基準とする言語（日本語，あるいは，英語）を定めてから，類似度を計算する必要があった．

本稿では，研究者マッチングの評価として，科研費の共同研究者という観点からの評価 (評価 1)，および，同じ分野の研究者という観点の評価 (評価 2) について検討した．

今後は，これらの課題を踏まえて，研究者マッチングの研究を進める予定である．

**謝辞**

本研究は JSPS 科研費 21H03769，および，電気通信普及財団の助成を受けたものである．

**参考文献**

[1] 吉田光男. 特集：「研究評価と学術情報分析」特集「研究評価と学術情報分析」にあたって. 人工知能, Vol. 38, No. 3, pp. 358–359, 2023.

[2] 大向一輝. 学術情報サービスのメタデータ・デザイン (特集メタデータの現在). 情報の科学と技術, Vol. 60, No. 12, pp. 495–500, 2010.

[3] 相澤彰子. 自然言語処理による学術論文の解析：現状と展望. 人工知能, Vol. 38, No. 3, pp. 375–383, 05 2023.

[4] 桂井麻里衣. 学術データに基づく研究者の特徴表現. 人工知能, Vol. 38, No. 3, pp. 392–398, 2023.

[5] Jing Li, Feng Xia, Wei Wang, Zhen Chen, Nana Yaw Asabere, and Huizhen Jiang. Acrec: a co-authorship based random walk model for academic collaboration recommendation. In Chin-Wan Chung, Andrei Z. Broder, Kyuseok Shim, and Torsten Suel, editors, *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pp. 1209–1214. ACM, 2014.

[6] 林容央, 桂井麻里衣. 研究者の活動可視化に向けたウェブページの多クラス分類. 人工知能学会全国大会論文集 第 37 回全国大会 (2023), pp. 2P4–GS–11–02, 6 2023.

[7] Tom Hope, Doug Downey, Oren Etzioni, Daniel S. Weld, and Eric Horvitz. A computational inflection for scientific discovery, 2022.

[8] 榊剛史, 松尾豊, 市瀬龍太郎, 武田英明, 石塚満. 論文データベースからの研究トピック抽出. 人工知能学会全国大会論文集, No. 0, pp. 43–43, 2005.

[9] 小林雄太, 松本裕治ほか. 構成要素を考慮した論文の分散表現に基づく類似論文検索. 研究報告自然言語処理 (NL), Vol. 2017, No. 20, pp. 1–6, 2017.

[10] Yuto Chikazawaand Marie Katsurai and Ikki Ohmukai. Multilingual author matching across different academic databases: a case study on kaken.