複数の異なる学術情報データベースを対象とした著者同定支援システムに関する検討

桂井麻里衣† 大向 一輝††

† 同志社大学理工学部 〒610–0394 京都府京田辺市多々羅都谷 1-3
†† 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@mm.doshisha.ac.jp, ††i2k@nii.ac.jp

あらまし 研究者の専門分野の把握や研究機関のパフォーマンス分析には，各研究者と過去の業績情報を正確に対応付ける必要がある．全ての成果を一元管理する学術情報データベースは国内外に存在しないため，複数の異なるデータベースからの情報集約が求められる．しかし，複数データベースを横断した研究者 ID は整備されていないため，自動的な名寄せは困難である．そこで本研究では，異なる種類の学術情報データベースを対象とした著者同定支援システムの初期検討として，同一人物候補のランキング手法を提案する．提案手法では，様々なデータベースで利用可能なメタデータに基づき複数の類似度を定義し，それらの値の総和を同一人物らしさのスコアとみなす．本文の最後には，CiNii Dissertations 中の博士論文と KAKEN 研究者 ID における重複漢字氏名を対象とした著者同定実験を行い，提案手法の有効性について検討する．

キーワード 学術情報データベース，著者同定，研究者分析，名寄せ

1. はじめに

科学技術の動向分析や研究企画戦略の策定において，研究者の専門分野の把握や研究機関のパフォーマンス分析は必要不可欠である．例として，研究者が出版した論文の被引用数や掲載誌のインパクトファクターは，アウトプットを客観的に評価する指標として頻繁に用いられる [1]．加えて，各研究組織に属する研究者情報を整理し，共同研究推進に活用する動きも見られる [2]．著者らのこれまでの研究では，国内のあらゆる分野の研究者を対象としたプロファイリング [3] や共同研究者推薦 [4]，科研費獲得における組織内共同研究関係の分析手法 [5] を提案した．このような研究支援を目的とした研究には，いずれも各研究者と過去の業績情報を正確に紐付けたデータベースが必須となる．現在，国内には学術論文を収録した CiNii Articles(注1) や科研費採択課題を収録した KAKEN(注2)，博士論文を収録した CiNii Dissertations(注3) など様々な学術情報データベースが存在する．しかしながら，システム横断的な研究者 ID が整備されておらず，各データベースに同一人物に関する記述の存在が保証されていないことから，複数データベースにおける自動的な名寄せは困難である．同姓同名研究者の存在はデータベース管理における大きな問題の一つとして知られており，様々な著者同定手法が提案されてきた [6]．しかしながら，複数データベースから異なる種類の学術情報を自動集約する技術は未だ提案されていない．

そこで本論文では，複数の異なる学術情報データベースにおける著者同定支援システムの初期検討として，同一人物候補のランキング手法を提案する．はじめに，二つのデータベース間で著者名の文字列を比較し，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）に着目し，異なる観点から類似度を定義する．複数類似度の総和からなるスコアに基づき，同一人物らしいと考えられる順に同一人物候補をソートする．本文の最後には，KAKEN の研究者 ID を CiNii Dissertations の博士論文著者に紐付ける著者同定実験により提案手法の有効性を評価する．

以降，2 章で本研究の関連研究を説明し，3 章でメタデータの類似度に基づく同一人物候補のランキング手法を提案する．4 章では KAKEN および CiNii Dissertations における同姓同名著者人数の詳細を示したあと，提案手法の性能を評価するための実験を行う．最後に，5 章において本文をまとめ，今後の研究課題について述べる．

2. 関連研究

本章では，本研究の関連研究として，同一データベースにおける著者同定手法と，複数データベース間の著者同定についてそれぞれ述べる．

2.1 同一データベース内の著者同定手法

研究者のプロファイリング [2,3] や業績評価には，研究者と過去の業績情報（学術論文や研究助成課題など）が正確に対応付いている必要がある．これを実現するには，研究者に固有の ID を割り当てることが有効である [7]．Thomson Reuters

(注1)：https://ci.nii.ac.jp/
(注2)：https://kaken.nii.ac.jp/
(注3)：https://ci.nii.ac.jp/d/

は ResearcherID(注4)，Elsevier は Author Identifier(注5)をそれぞれ提供している．国内の CiNii Articles では，科研の報告書に記載されたことのある論文を KAKEN 研究者 ID に紐付けることで正確な業績管理を進めている [3].

与えられた学術論文を既存の著者に割り当てる著者同定手法も盛んに研究されており，教師あり・半教師あり学習または教師なしアプローチに大別される．例として，Han ら [8] は人手で著者 ID を付与した論文集合を教師データとし，論文タイトルの単語や共著者名を特徴としてナイーブベイズと SVM を学習した．一方，代表的な教師なしアプローチでは，著者氏名や共著者氏名，所属，参考文献などのメタデータに基づき論文間のペアワイズ類似度を算出し，階層的クラスタリングにより同一著者のものと考えられる論文集合を得る [9]．これらの手法で用いるメタデータのうち，参考文献は分野を特徴付ける重要な属性として知られている．しかし，参考文献は入手コストが高く，かつ論文以外の学術情報ではその属性をもたないという問題がある．共著者氏名のマッチングに関しても，単著による学術情報では適用が不可能である．そこで本論文では，多くの学術情報データベースで保有しうるメタデータのみに着目し，比較対象のいずれかが単著でも利用可能な類似度指標を提案する．

2.2 異なるデータベースを横断した著者同定

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) [10] という組織が世界中の研究者に対し著者 ID を発行している．しかし，研究者自らが ID を取得し，あらゆる学術情報データベースで ORCID を手動登録することが求められているのが現状であり，未だ学術情報の一元管理は実現していない．

学術情報データベースを横断した著者同定に関しては，同一データベース内での手法に比べて非常に報告が少ない．文献 [11] では，ウェブ上のテキスト文書に出現する氏名と論文著者を紐付けるための確率モデルが提案されている．この手法は，テキスト文書内の文脈から氏名を分類することを想定している．一方，全ての学術情報データベースで長文テキストの入手が保証されておらず，本研究への直接的な応用は困難と考えられる．

3. CiNii Dissertations と KAKEN における同一人物候補のランキング手法

本論文では，複数の異なる学術情報データベースにおける著者同定支援システムの初期検討として，ある特定の問題設定を取り上げる（3.1 節）．提案手法では，まず二つのデータベースにおける氏名の文字列マッチングから同一人物候補集合を構築する（3.2 節）．次に，多くの学術情報で保有しうるメタデータを対象に，様々な観点から類似度を定義する（3.3 節）．複数の類似度の総和から同一人物らしさのスコアを算出し，スコアが上位となる研究者を提示することで，データベース管理者の作業の効率化を目指す．

(注4)：http://www.researcherid.com/
(注5)：http://www.elsevier.com/online-tools/scopus

3.1 問題設定

本論文では，博士論文データセット D_1 と科研費データセット D_2 における著者マッチング問題を考える．漢字氏名 α を著者にもつ博士論文集合と KAKEN 研究者 ID 集合をそれぞれ $\Omega_1(\alpha)$, $\Omega_2(\alpha)$ とし，博士論文 d（研究者 ID i）の所属文字列を $\mathrm{aff}(d)$ ($\mathrm{aff}(i)$) で表す．以降，集合 Ω の要素数を $|\Omega|$ で表す．

一般に，各博士論文 $d \in \Omega_1(\alpha)$ は以下のメタデータをもつ．

- 著者の漢字氏名（α に相当）.
- 論文タイトル：$\mathrm{title}(d)$.
- 著者の所属（学位授与機関）：$\mathrm{aff}(d)$.
- 論文発表年：$y(d)$.

これらは学術論文の書誌情報を管理するデータベースの多くが保有すると考えられる．

一方，各 KAKEN 研究者 ID $i \in \Omega_2(\alpha)$ は以下のメタデータをもつ．

- 研究者の漢字氏名（α に相当）.
- 研究者のこれまでの所属文字列集合：$A(i)$.
- 過去に採択された科研課題のタイトル集合：$T(i)$.
- 過去に採択された科研の共同研究者 ID 集合：$Co(i)$.
- 過去に採択された科研の共同研究者の所属文字列集合：$CoA(i) = \{\mathrm{aff}(j) | j \in Co(i)\}$.
- 過去に採択された科研の研究開始年集合：$Y(i)$.

これらのメタデータについても，研究者 ID が一部整備されている学術情報データベースであれば容易に入手可能といえる．

3.2 氏名文字列マッチングによる同一人物候補の抽出

漢字氏名 α に該当する著者集合 $\Omega_1(\alpha)$，研究者 ID 集合 $\Omega_2(\alpha)$ を構築するために，データセット中の氏名文字列が完全に一致する人物のみを抽出する．本論文では日本語の学術情報を想定しているため，著者/研究者氏名は漢字で比較する．読み表記は全てのデータベースに存在する保証がないため利用しない．なお，英語論文の場合はファーストネームのイニシャル表記が起こりうるため，異なる言語の学術情報データベースを対象とする場合は本ステップを改良する必要がある．

3.3 メタデータに基づく類似度算出

博士論文 $d \in \Omega_1(\alpha)$ と KAKEN 研究者 ID $i \in \Omega_2(\alpha)$ のペアについて，学術情報データベースから容易に収集可能なメタデータとして，(A) 著者の所属文字列類似度，(B) 共同研究者の所属文字列類似度，(C) 研究課題タイトル，(D) 研究発表年に基づく類似度算出方法を提案する．

A）著者の所属文字列類似度.「○○大学□□学部」や「△△研究所××系」という表記のうち，部局情報を除去した所属文字列である○○，△△のみを用いて次式を算出する．

$$Sim_{\mathrm{aff}}(d,i) = \max_{z \in A(i)} StringSim(\mathrm{aff}(d), z), \qquad (1)$$

$$StringSim(x,y) = \frac{1}{2}\Big\{(1 - LD(x,y))) + JC(x,y))\Big\}.$$

上式において，$LD(x,y)$ は二つの文字列 x, y の最大文字数で正規化した文字列間 Levenshtein 距離，$JC(x,y)$ は文字列 x, y 間の Jaccard 係数を表す．研究者の所属は変わりうるため，文字列類似度の最大値をとることで同一機関または周辺地域に存

在していた経歴を発見するねらいである．

B) 共同研究者の所属文字列類似度．従来の著者同定問題において，共著者氏名は識別力の高い特徴となることが知られている．しかし，博士論文のように単著で執筆された文献については，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する．具体的に本問題設定では，博士論文 d の著者の所属と，KAKEN 研究者 ID i の過去の共同研究者の所属に着目し，次式の類似度を算出する．

$$Sim_{\text{coaff}}(d,i) = \frac{1}{|CA(i)|} \sum_{z \in CA(i)} Sim_{\text{aff}}(d,z), \qquad (2)$$

これにより，著者本人の所属が変わっていたとしても，以前の研究機関に科研の共同研究者を残している場合を考慮できる．

C) 研究課題タイトル．研究課題のタイトルで用いられている単語集合から研究内容の類似性を評価する．まず，日本語 Wikipedia のエントリを辞書に登録した形態素解析エンジン MeCab(注6) を用いてタイトルから名詞集合を抽出する．このとき，学術用語の過分割を避けるため，名詞が連続する場合にはバイグラムも抽出する．$\Omega_1(\alpha)$, $\Omega_2(\alpha)$ における全研究課題の名詞集合からボキャブラリを構築し，$d \in \Omega_1$，$i \in \Omega_2$ に対し term frequency (TF) ベクトル $\boldsymbol{w}(d)$, $\boldsymbol{w}(i)$ を算出する．$\boldsymbol{w}(i)$ の算出には，研究課題タイトル集合 $T(i)$ 全ての単語を用いる．ベクトル間 $\boldsymbol{w}(d), \boldsymbol{w}(i)$ のコサイン類似度を研究課題タイトル類似度 $Sim_{\text{title}}(d,i)$ とする．

D) 発表年．研究者として活動している期間が近いほど同一人物の可能性が高いといえる．そこで，二つのデータベース内での研究発表年に基づき次式を算出する．

$$Sim_{\text{year}}(d,i) = \frac{1}{N} \min_{y' \in Y(i)} |y(d) - y'| \qquad (3)$$

正規化定数 N は想定しうる最長研究年であり，本稿では $N = 30$ と設定する．

3.4 著者候補のランキング

前節で提案した複数の類似度に基づき同一人物候補を総合的にランキングするために，次式のスコアを算出する．

$$Sim(d,i) = Sim_{\text{aff}}(d,i) + Sim_{\text{coaff}}(d,i) + Sim_{\text{year}}(d,i) + Sim_{\text{title}}(d,i) \qquad (4)$$

最終的に，対象とする KAKEN 研究者 ID $i \in \Omega_2(\alpha)$ に対し，最も同一人物の著作物らしい博士論文 $d^* = \arg\max_{d \in \Omega_1(\alpha)} Sim(d,i)$ を提示する．

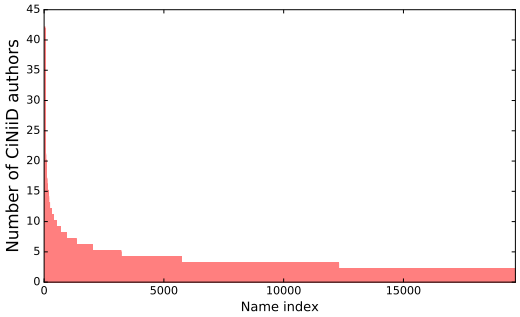
4. 実　　験

本章では，CiNii Dissertations および KAKEN 研究者 ID における実際の同性同名人数分布を示したあと，出現頻度が上位となる漢字氏名に対する著者同定実験結果を報告する．

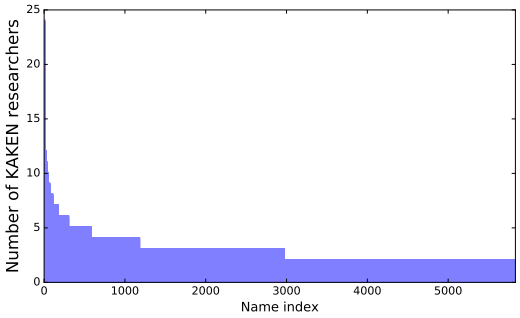
(注6)：http://taku910.github.io/mecab/

表 1　博士論文データセット D_1 と科研データセット D_2 の詳細．

D_1, D_2 の両方に存在するユニークな漢字氏名の数	159,657
D_1 の博士論文数	207,215
D_1 内で同姓同名が存在する漢字氏名の数	39,228
D_2 の研究者 ID 数	176,544
D_2 内の科研課題数	138,894
D_2 内で同姓同名研究者が存在する漢字氏名の数	11,665



(a) 博士論文データセット D_1 の著者氏名と人数のヒストグラム．



(b) 科研データセット D_2 の研究者氏名と人数のヒストグラム．

図 1　博士論文データセット D_1 および科研データセット D_2 における氏名と人数のヒストグラム．紙面の都合上，横軸の範囲は同姓同名をもつ氏名数の半数に設定した．

表 2　データセット中の論文・科研タイトルにおける上位頻出単語．

研究 的研究 細胞 開発 解析 機能 機構 構造 解明 学的 影響 制御 検討 応用 遺伝子 評価 分子 システム 効果 基礎 組織 実験 形成 反応 治療 神経 モデル 発現 作用 特性

4.1 CiNii Dissertations および KAKEN 研究者 ID の漢字氏名重複状況

2017 年 12 月 25 日現在 CiNii Dissertations に収録されている約 60 万本の博士論文から 558,280 個のユニークな著者漢字氏名を抽出した．各漢字氏名をクエリとして KAKEN 研究者 ID を検索した結果，KAKEN 内に存在した氏名数は 159,657 であった．このように重複した氏名を再度クエリとして用いて，博士論文データセット D_1，科研データセット D_2 をそれぞれ構築した．各データセットの詳細を表 1 に示す．また，各データセットの氏名－研究者人数分布の一部を図 1 に示す．重複漢字氏名のうち，多くは研究者が二名のみ該当する状況であるが，最頻出氏名は D_1 に 42 名，D_2 に 24 名存在した．

3.3 節で提案した研究課題タイトルの類似度指標算出にあたり，全文書の 2% 以上に出現した単語を表 2 に示す．これらの

表 3　評価用データセットの詳細.

氏名 α の番号	$D_1(\alpha)$	$D_2(\alpha)$	正解ペア数
1	42	24	20
2	35	22	18
3	33	17	14
4	16	14	10
5	23	13	10
6	19	13	8
計	168	103	80

表 4　類似度尺度の組合せで算出した類似度上位 K 人を提示したときの正解数.

	$K=1$		$K=3$		$K=5$	
順位	組合せ	正解数	組合せ	正解数	組合せ	正解数
1	ABCD	41	ABCD	52	ABCD	58
2	ABC	40	ACD	52	ACD	57
3	ACD	39	ABD	51	ABD	56
4	BCD	39	AC	50	AC	55
5	AC	38	ABC	49	ABC	54
6	ACD	37	AD	47	BCD	54
7	AD	36	BCD	47	AD	52
8	BC	35	BC	45	BC	51
9	BD	33	AB	43	AB	50
10	AB	29	BD	42	BD	48
11	A	26	A	40	A	43
12	C	25	B	33	CD	40
13	B	21	C	32	B	38
14	CD	19	CD	30	C	32
15	D	7	D	21	D	31

単語は著者を特徴付ける効果が低いと考えられるためストップワードとした.

4.2　提案手法による同一著者候補ランキング結果

本節では，提案手法の性能を検証するために，KAKEN 研究者 ID に同一著者の博士論文を提示する実験を行う．全ての同姓同名研究者に対し正解データを用意することは非常に困難であったため，図 1(b) において出現頻度が上位 6 個となる漢字氏名のみに対し本論文の第一著者が手動で正解データを与えた．構築した評価用データセットの詳細を表 3 に示す．なお，表中の正解ペア数とは「著者が確証をもって正しいといえる KAKEN 研究者 ID と博士論文のペア数」をさす．博士号をもたない研究者は正解ペアから除外した．正解データ構築には各研究者の経歴調査と確認に多大な労力と時間を要し，人的コストと難易度の両方が極めて高いタスクといえる．

各漢字氏名 α について，正解ペアに存在する KAKEN 研究者 ID に対し，博士論文著者候補 $\Omega_2(\alpha)$ のうち式 (4) が最大となる上位 K 個の博士論文を選出した．選出した K 名に本人の博士論文が含まれていれば正解とカウントした．式 (4) において，類似度 A〜D それぞれの効果を分析するため，利用しない類似度を 0 とすることで組合せを変更した．$K=1,3,5$ の場合について得られた結果を表 4 に示す．表より，いずれの K の値についても，全ての類似度の和を用いたときに最も正解数が多いことがわかる．類似度を単独で用いる場合は著者本人の所属が最も強力な特徴となった．また，研究年のみに着目すると正しい著者同定は困難だが，他の特徴と組み合わせることで同定の信頼性を高められることがわかった．しかし，全ての類似度を用いた場合でも未だ正解率は高いとはいえない．本論文では類似度指標の総和という簡便な方法から最終的なスコアを算出したが，今後は各類似度指標の適切な重みを決定する手法が求められる．また，提案した研究課題タイトルの類似度は使用単語の重複から算出しているため，関連性を見落とす可能性がある．したがって，今後は同義語や関連語などの事前知識を導入し，研究課題間の関連性を適切に捉える方法を導入する予定である.

5. ま と め

本論文では，複数の異なる学術情報データベースを対象とした著者同定支援システムの初期検討として，メタデータ類似度に基づく同一人物候補のランキング手法を提案した．はじめに，データベース間で著者名の文字列比較に基づくマッチングを行い，同一著者候補集合を構築した．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）に着目し，複数の類似度指標を提案した．本論文では，これらの指標を集約する簡便な方法として，値の総和を最終的なスコアとみなして著者候補をランキングした．本文の最後には，提案手法の有効性を評価するために，KAKEN の研究者 ID と同一著者による CiNii Dissertations 内の博士論文を選出する実験を行った．提案した全ての類似度を用いることで最大の正解数を達成したが，指標の重み付け方法については検討の余地があるといえる．これまでに著者らが提案した研究者プロファイリング手法 [3] では，論文概要をトピックベクトルに変換することで単語集合から潜在的な分野を抽出した．今後は，このような分野推定手法を導入することで，研究課題の類似性を適切に捉える必要がある．実用化にあたっては，提案手法は同一人物候補を提示するのみであり，最終的な著者同定は人手による確認作業を想定している．したがって，確認内容を正解データとしてフィードバックすることでランキングを改良する仕組みも検討する予定である.

文　献

[1] M. Kotsemir and S. Shashnov. Measuring, analysis and visualization of research capacity of university at the level of departments and staff members. *Scientometrics*, Vol. 112, No. 3, pp. 1659–1689, Sep 2017.

[2] 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.

[3] M. Katsurai, I. Ohmukai, and H. Takeda. Topic representation of researcher's interests in a large-scale academic database and its application to author disambiguation. *IEICE Trans. Information and Systems*, Vol. E99-D, No. 4, pp. 1010–1018, April 2016.

[4] M. Araki, M. Katsurai, I. Ohmukai, and H. Takeda. Interdisciplinary collaborator recommendation based on research content similarity. *IEICE Trans. Information and Systems*,

Vol. E99-D, No. 4, April 2017.

[5] 荒木将貴, 桂井麻里衣, 大向一輝, 武田英明. 大学における部局横断型共同研究の活発さを把握する指標の検討. 日本データベース学会和文論文誌, Vol. 16-J, No. 2, 2018.

[6] 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.

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

[8] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis. Two supervised learning approaches for name disambiguation in author citations. In *Proc. Joint ACM/IEEE Conf. Digital Libraries (JCDL)*, pp. 296–305, 2004.

[9] H. Wu, B. Li, Y. Pei, and J He. Unsupervised author disambiguation using Dempster–Shafer theory. *Scientometrics*, Vol. 101, No. 3, pp. 1955–1972, Dec 2014.

[10] 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.

[11] W. Shen, J. Han, and J. Wang. A probabilistic model for linking named entities in web text with heterogeneous information networks. In *Proc. ACM SIGMOD Int. Conf. Management of Data (SIGMOD)*, pp. 1199–1210, 2014.