The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

研究内容の時間変化と所属情報を考慮した類似研究者検索に関する検討

A Note on Similar Researcher Retrieval Considering Temporal Changes of Research Content and Affiliations

西澤 浩之 *1 Hiroyuki Nishizawa　桂井 麻里衣 *2 Marie Katsurai　大向 一輝 *3 Ikki Ohmukai　武田 英明 *3 Hideaki Takeda

*1同志社大学大学院理工学研究科 Graduate School of Science and Engineering, Doshisha University

*2同志社大学理工学部 Faculty of Science and Engineering, Doshisha University

*3国立情報学研究所 National Institute of Informatics

As a research theme becomes more complicated, the range of knowledge necessary for research spreads wider. The collaborative research by multiple researchers is known to be effective for doing complicated research. Existing methods on research collaborator recommendation calculate the relevance between researchers using research content similarity and co-authorship relationships. This paper focuses on introducing new features, i.e., *temporal changes of research content* and *affiliations*, into researcher profiling for collaborator recommendation. In the proposed method, we use the titles and abstracts of academic papers as textual information for modeling research topics. Specifically, calculating a topic vector of a researcher's publications in each year, we represent the researcher's interests using a series of topic vectors. We also obtain each researcher's affiliation information from the Database of Grants-in-Aid for Scientific Researcher, named KAKEN. Based on the resulting profiles consisting of temporal changes of research content and affiliations, we present a novel similarity measure of researchers. In experiments, we present topic extraction results from a subset of CiNii Articles and show an interface example of similar researcher retrieval based on the proposed method.

1. はじめに

研究課題が複雑化するにつれて，研究に必要な知識範囲は広がる．複雑化した研究課題を円滑に進めるには，複数の研究者が共同で研究に取り組むことが有効といわれている．共同研究と生産性の関係や共同研究の推進の効果については多数の議論がある．例として，文献 [Abramo 09] は，共同研究と生産性の関係を実証するために，異分野共同研究や外部との共同研究，産学連携とその研究成果の相関を調べた．文献 [Wang 17] は，共同研究者の推薦が学術研究の発展を促進することを証明した．以上の共同研究の重要性から，効率的に共同研究者候補を検索する手法が種々提案されている．我々の過去の研究では，研究者の論文集合から多次元トピックベクトルを抽出し，トピックベクトル間の類似度に基づき研究者間の類似度を算出した [Araki 17]．この手法では，学術論文の出版日時を考慮しておらず，各研究者に単一のトピックベクトルのみを割り当てる．一方，学術情報に限定しない情報推薦の研究では，ユーザのプロファイルを構築する際に，ユーザ嗜好の時間変化を考慮すべきといわれている [Liang 12]．加えて，研究者間のコミュニケーションを促進するためには，研究内容の類似度の他にも共通点を発見しうる新たな特徴を追加すべきと考えられる．

そこで本研究では，研究内容の時間変化と所属情報を考慮した類似研究者検索手法を提案する．提案手法は，(i) 学術論文集合と所属情報に基づく研究者プロファイリングと (ii) 研究者プロファイルに基づく類似研究者検索から構成される．はじめに，年ごとに分割した学術論文集合のタイトル・概要から各研究者の研究内容特徴ベクトルを算出するとともに，研究者の所属とその位置情報をウェブから取得する（2 章）．構築したプロファイルに基づき，研究者間類似度を定義する（3 章）．これにより，研究内容の変遷が類似しており，かつ地理的に近い研究者が検索可能となる．本文の最後には，CiNii Articles*1 の論文を用いた研究トピック抽出結果と，提案手法に基づく研究者検索インタフェースの構築例を示す．

連絡先: 西澤浩之，同志社大学大学院理工学研究科，〒 610-0394 京田辺市多々羅都谷 1-3，nishizawa@mm.doshisha.ac.jp

2. 学術論文集合と所属情報に基づく研究者プロファイリング

本章では，学術論文集合と所属情報に基づく研究者プロファイリング手法を説明する．以降，学術論文集合から年ごとに研究内容を抽出する方法と所属の位置情報を算出する方法について説明する．

2.1 学術論文からの研究内容抽出

本節では，学術論文集合から年ごとに研究内容を抽出する方法を説明する．学術論文の本文から研究内容に直接関連する文章のみを抽出することは困難であるため，各論文のタイトルと要約を一つの文書とみなし，研究内容の分析に用いる．手法の概要を図 1 に示す．具体的には，各研究者の論文を出版年ごとに分割し，以下の研究内容抽出手順を適用する．

1. 論文テキストから名詞のみを抽出し，Bag-of-Words (BoW) 表現に変換する．
2. BoW 表現に対して Latent Dirichlet Allocation (LDA)[Blei 03] を適用し，トピックベクトルを算出する．
3. 閾値以下のベクトル要素を 0 にし，L2 正規化を行う．

以降，各手順の詳細を説明する．

2.1.1 テキストからの BoW 算出

まず，LDA の学習と推定の精度向上を目的とし，以下のように学術論文のタイトル・要約テキストを前処理する．

*1 https://ci.nii.ac.jp/

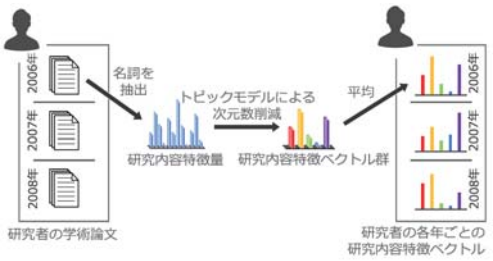

図 1: 研究者の研究内容特徴ベクトル群算出の概要図.

- 記号の除去.
- 数字の除去.
- 全角空白の除去.
- 括弧の除去.
- URL 文字列の除去.
- 半角カタカナを全角カタカナに変換.
- 全角英字を半角英字に変換.
- 英大文字を英小文字に変換.

次に，日本語形態素解析エンジン MeCab[*2] を用いてテキストから名詞のみを抽出する．得られるボキャブラリには，「研究」「提案」「一般」のように研究者の研究内容を特徴付ける効果が低いと考えられる単語が混在する．そこで，出現頻度が最上位または最下位となる語をストップワードとして除去する．以上により得られた単語リストを用いて，各論文に対し BoW ベクトルを算出する．

2.1.2 確率的トピックモデルによる次元削減

BoW の次元数はデータセット中に登場するユニークな単語数に相当する．語の関連性を考慮し，かつ類似度算出時の計算量を削減するために，LDA を適用して次元数を削減する．LDA は確率的トピックモデルの一つであり，文書は複数のトピックから構成されていると仮定する．文書集合をトピックに分解し，各文書をトピックの組み合わせで表す．一連の文書に用いる場合，トピックは集合のテーマとして解釈可能であり，文書のトピック分布は各文書がどのテーマに関するものかを示す．これにより，BoW の次元をトピック数 K まで削減する．

具体的には，まずデータセットの全論文を用いて LDA モデルを学習し，トピック分布を算出する．次に，研究者 a の論文集合を年 y ごとに分割する．分割された論文集合 $S_{a,y}$ ごとに，学習したモデルで各論文 $d \in S_{a,y}$ のトピックベクトル $\boldsymbol{\theta}_d$ を算出する．最後に，次式のように集合 $S_{a,y}$ のトピックベクトルの平均を算出する．

$$\boldsymbol{v}_{a,y} = \frac{1}{|S_{a,y}|} \sum_{d \in S_{a,y}} \boldsymbol{\theta}_d \tag{1}$$

得られた $\boldsymbol{v}_{a,y}$ を研究者 a の年 y における研究内容特徴ベクトルとみなす．

2.1.3 研究者の研究内容の抽出

算出した研究内容特徴ベクトル $\boldsymbol{v}_{a,y}$ の要素に着目したとき，あるトピックに非常に小さな値が割り当てられている可能性がある．そのようなトピックは研究者の研究内容を表現する際に重要ではないと考えられる．そこで，ベクトル $\boldsymbol{v}_{a,y}$ の k 番目のトピックに対する要素 $v_{a,y,k}$ に対し，次式の閾値処理を適用する．

$$v_{a,y,k} = \begin{cases} v_{a,y,k} & (v_{a,y,k} \geq Threshold) \\ 0 & (v_{a,y,k} < Threshold) \end{cases} \tag{2}$$

これにより，研究者の主要な研究内容を表すトピックのみが残される．最後に，ベクトル $\boldsymbol{v}_{a,y}$ を L2 正規化する．

2.2 研究者の所属情報の取得

本節では，各研究者の所属の位置情報を算出する方法を説明する．はじめに，KAKEN[*3] の研究者詳細画面から各研究者の所属文字列を抽出する．研究者詳細画面は一意の研究者番号を含む URL で管理されているため直接アクセス可能であり，研究者の氏名，研究者番号，所属，研究分野，キーワード，研究課題が記載されている．KAKEN を用いることで，各大学・研究所の研究者ページへ個々にアクセスすることなく研究者のメタデータが得られる．得られた所属情報から Google Maps API[*4] を用いて位置情報（緯度・経度）を得る．Google Maps API は地図の描画や緯度経度の検索，ルート検索などの機能があり，その一部を研究者検索インタフェースの構築に利用する．

3. 研究者プロファイルに基づく類似研究者検索

本章では，前章で構築した研究者プロファイルに基づく類似研究者検索手法を提案する．まず，研究者 a, b に対し，研究内容の時間変化の類似度 $ResearchContentSim(a, b)$ を次式により算出する．

$$ResearchContentSim\ (a, b) = \frac{1}{n} \sum_{y=y_0}^{y_n} CosSim\ (\boldsymbol{v}_{a,y}, \boldsymbol{v}_{b,y}) \times \log\ (y - y_0 + 2) \tag{3}$$

$$CosSim\ (\boldsymbol{v}_a, \boldsymbol{v}_b) = \frac{\sum_{t=1}^{k} v_{a,t} \times v_{b,t}}{\sqrt{\sum_{t=1}^{k} v_{a,t}^2} \times \sqrt{\sum_{t=1}^{k} v_{b,t}^2}} \tag{4}$$

上式において，n は研究内容特徴ベクトル群の分割年の総数を示す．log 関数を適用することで，最新の研究ほど類似度に寄与する．次に，位置情報による研究者間の類似度 $LocateSim(a, b)$ として，Google Maps API を用いて取得された緯度経度から計算された直線距離を正規化し類似度へ変換する．

以上の二つの類似度を組み合わせ，研究者 a，b 間の最終的な類似度を次式で算出する．

$$Sim\ (a, b) = \phi \times ResearchContentSim\ (a, b) + (1 - \phi) \times LocateSim\ (a, b) \tag{5}$$

上式において，ϕ は二つの類似度のバランスを調整するためのパラメータである．与えられた研究者 a に対し，式 (5) の類似度が上位となる研究者 b を提示することで，研究者 a の類似研究者を容易に把握できる．

4. 実験

4.1 データセット

CiNii Articles からの論文収集にあたり，2003 年から 2010 年の間に電子情報通信学会 [*5] で一度でも発表したことがあり，

*2 http://taku910.github.io/mecab/

*3 https://kaken.nii.ac.jp/

*4 https://developers.google.com/maps/

*5 http://www.ieice.org/jpn/

表 1: 実験用のデータセットの詳細.

研究者数	4,680 名
学術論文数	38,098 本
論文出版年	2003-2010

表 2: LDA によって推定された研究トピックの例．トピックの解釈を見出しで示す.

トピック 24 テキストマイニング		トピック 50 検索エンジン		トピック 96 画像処理	
単語	確率	単語	確率	単語	確率
単語	0.092	検索	0.254	画像	0.428
クラスタリング	0.045	分類	0.118	処理	0.037
辞書	0.041	類似	0.095	復元	0.026
出現	0.039	属性	0.040	エッジ	0.025
頻度	0.037	クエリ	0.030	局所	0.022
語彙	0.025	精度	0.020	画素	0.019
日本語	0.020	エンジン	0.018	テクスチャ	0.019
漢字	0.018	キーワード	0.018	カラー	0.014
対象	0.015	方法	0.014	解像度	0.014
共起	0.014	向上	0.012	特徴	0.014

トピック 112 推薦システム		トピック 133 VR・AR・MR		トピック 150 人工知能	
単語	確率	単語	確率	単語	確率
ユーザ	0.435	空間	0.395	学習	0.310
推薦	0.040	仮想	0.109	ルール	0.042
表情	0.038	現実	0.058	識別	0.039
興味	0.023	力	0.036	サンプル	0.028
嗜好	0.021	再現	0.034	問題	0.023
提示	0.021	世界	0.022	推論	0.020
サイト	0.017	覚	0.019	獲得	0.019
顔	0.016	環境	0.018	ニューラルネットワーク	0.019
意図	0.012	物理	0.017	入力	0.014
コンテキスト	0.010	複合	0.012	ファジィ	0.014

かつ KAKEN の ID をもつ研究者リストを構築した．実験用のデータセットの詳細を表 1 に示す．なお，所属の位置情報を取得する際，KAKEN のメタデータに明らかに誤った記載があったため，Google Maps API で位置情報が取得できる形式となるよう文字列を手動で修正した.

4.2 研究者の研究内容変化の特徴抽出

研究者の研究内容特徴ベクトルを抽出するために，全ての論文テキストに MeCab[*6] を適用して単語リストを生成した．MeCab の辞書には新語・固有名詞を考慮可能な mecab-ipadic-NEologd[*7] を使用した．その後ストップワードとしてデータセットの 20 %（7619 本）以上に出現する単語および 9 本以下にのみ出現する単語を除去した．LDA のトピック数は $K = 150$，ハイパーパラメータは $\alpha = \frac{50}{K}$，$\beta = 0.01$ と設定した．LDA によって推定された研究トピックとその単語分布の例を表 2 に示す．LDA により，6907 次元であった BoW が 150 次元まで削減され，低次元の研究内容特徴ベクトルが得られた．また研究内容特徴ベクトルから研究内容を抽出するにあたり，弱いトピックの判断基準として 0.1 を閾値とした.

4.3 類似研究者検索インタフェースの構築

提案手法に基づき，対象とする研究者の類似研究者を検索できる Web アプリケーションを構築した．本実験では，研究内容の類似性を重視して $\phi = 0.65$ とした．類似研究者検索インタフェースの画面表示例を図 2 に示す．対象研究者の氏名，研究者 ID を画面上部，所属と過去の論文リストを左部に示す．また，所属の位置情報は右部に地図表示する．画面右下には，提案手法において類似度が上位となった研究者を表示する．クリックされた研究者が新たに検索クエリとなるため，類似研究者を次々と検索可能となる.

*6 http://taku910.github.io/mecab/

*7 https://github.com/neologd/mecab-ipadic-neologd

図 2: 類似研究者検索インタフェースの画面例.

5. まとめと今後の展望

本研究では，共同研究者推薦の実現に向け，新たに研究内容の時間変化と所属情報を考慮した研究者のプロファイリングとそれに基づく類似度算出方法を提案した．実験では，提案手法の応用例として，類似研究者検索インタフェースを構築した．今後の課題として，研究者の所属の時間変化など，より多くの特徴をプロファイルに加えることが挙げられる．本稿では研究者の所属情報から Google Maps API を用いて緯度・経度を取得したが，同じ大学でもキャンパスごとに位置情報が異なる問題があった．KAKEN から取得した情報ではキャンパス情報が収録されていないため，研究内容と所属情報に基づく在籍キャンパスの推定も一つの課題となりうる．今後の展望として，複数分野の研究者に提案手法を適用することによる異分野共同研究者の推薦や，ユーザがより使いやすい研究者検索インタフェースの構築，各地方や学研都市の研究内容可視化インタフェースの構築を検討している.

参考文献

[Abramo 09] Abramo, G., D'Angelo, C. A., and Costa, Di F.: Research collaboration and productivity: is there correlation?, *Higher Education*, pp. 155–171 (2009)

[Araki 17] Araki, M., Katsurai, M., Ohmukai, I., and Takeda, H.: Interdisciplinary Collaborator Recommendation Based on Research Content Similarity, *IEICE Trans. Information and Systems*, pp. 785–792 (2017)

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Leraning Research*, pp. 993–1022 (2003)

[Liang 12] Liang, H., Xu, Y., Tjondronegoro, D., and Christen, P.: "Time-aware Topic Recommendation Based on Micro-blogs, in *Proc. the 21st ACM Int. Conf. Information and Knowledge Management*, p. 16571661 (2012)

[Wang 17] Wang, W., Cui, Z., Gao, T., Yu, S., Kong, X., and Xia, F.: Is Scientific Collaboration Sustainability Predictable?, in *Proc. the 26th Int. Conf. World Wide Web Companion*, pp. 853–854 (2017)