ISS-P-047

著者の貢献度に基づく 単一学術情報リポジトリにおける名寄せ手法の提案

村田 海優 † 楠 和馬 ‡ 寺本 優香 ‡ リュウ ショウウ ‡ 波多野 賢治 †
† 同志社大学文化情報学部 ‡ 同志社大学大学院文化情報学研究科

1 はじめに

近年，国内外を問わず，学術雑誌・会議の種類が多岐に渡っており，論文の本数が増加し続けている．これら学術論文情報から研究者個々人の情報収集に関する研究が行われている [1]．しかし，学術情報リポジトリ上では同姓同名研究者の判別不可能な場合が多く，これらの判別を人手で行うことは，多大な時間と労力を要する．

本研究では論文情報を基に同一著者か否か判定する方法を提案する．また，著者同一判定モデルに利用する新たな統計量として著者の貢献度を提案する．

2 関連研究

桂井らは，CiNii Dissertations[1] と KAKEN[2] の学術情報リポジトリを用いて，著者同一判定を行っている．これらリポジトリに含まれている著者の所属，題目などの論文情報を文字列の類似度で論文間の差異を定量化している．その各論文情報の類似度を総和することで同一著者である度合いを表すスコアとして提案している [2]．

文献 [2] の手法には以下の三つの問題があり，本研究ではこれら問題に対応した手法を提案する．

- 定量化した値を利用したスコアの算出式が恣意的である．
- 著者の所属情報を利用しており，所属情報の無いリポジトリに適応できない．
- 二種類のリポジトリの使用で情報を補っているため，他の学術情報リポジトリに利用できない．

3 提案手法

本研究では，多様な学術情報リポジトリに適用可能な同一著者か否かの判定方法の提案を行う．そのため，多くのリポジトリと共通する論文情報を持つ DBLP[3] データセットを使用する．DBLP では著名な著者を同姓同名の論文執筆者と区別するために，一部の著者名に識別番号が人手で付与されている．したがって，本研究では個人が特定されている著者およびそれら著者が執筆している論文情報を利用し，著者同一判定モデルの構築を行う．

本研究で提案する手法は桂井らと同様に，同姓同名の著者を含む 2 件の論文情報の差異を基にそれら著者が同一人物か否かの二値分類を行う．学習手法には，複数の分類木により多数決評価を行い汎化性能の向上を図ることができ，欠損値を含むデータに対しても頑健なランダムフォレストを用いる．同姓同名著者の論文ごとに全て組み合わせてペアを作成し，その著者が同一人物か否か二値で示すしたものを目的変数とし，その論文情報の差異を定量化したものを説明変数とし，本研究では共著者数の差やページ数の差など計 6 種類を使用する．

また，著者と論文両方から類似度を測ることができる統計量として著者の貢献度を提案する．論文の筆頭著者，第二著者といった共著者の並び順は論文への貢献度合いを表す意図を含む場合が多い，と報告されている [3]．著者順の情報を利用することで筆者の貢献度 $C_t(a_i)$ を式 (1) により定量化する．

$$C_t(a_i) = 1 - \frac{1}{N}(i-1) \tag{1}$$

ある論文 t の著者 a_i の貢献度は式 (1) 中の，N は著者総数，i は著者順である．

4 評価実験

評価実験では，提案した同一著者判定モデルの予測精度を検証するために，十分割交差検証法を行う．識別番号付き著者およびそれら著者が執筆している全論文情報を学習用データとテスト用データに分割する．その際，学習用データと評価用データに同じ著者名が混在しないように，著者名の種類を基準に分割を行うことで対応する．著者同一判定モデルの評価は，評価指標である再現率，適合率，F 値を用いて行う．また，モデル構築の際に用いた各変数の重要度を算出し，提案した変数の重要性を確認する．

5 おわりに

本研究ではランダムフォレストを用いて，論文情報を基に著者同一判定のための二値分類モデルを提案した．

今後の課題は，DBLP に含まれている論文情報を利用しただけであるため，論文情報に関連する情報を二次データと関連付けることで提案した著者同一判定モデルに利用する必要がある．具体的には，タイトルを基に分野情報の抽出や，引用文献関係に基づき他の論文との関係性などが挙げられる．

参考文献

[1] 桂井, 小野：“語の共起のバースト検出に基づく研究トレンドの可視化”, DEIM Forum 2017 論文集 (2017). G7-2.

[2] 桂井, 大向：“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”, DEIM Forum 2018 論文集 (2018). P9-6.

[3] Venkatraman: “Conventions of scientific authorship”, https://doi.org/10.1126/science.caredit.a1000039 (2010). 【2018/02/08 閲覧】.

[1] CiNii: https://ci.nii.ac.jp/d/【2019/02/08 閲覧】

[2] KAKEN: 科学研究費助成事業データベース https://kaken.nii.ac.jp/ja/【2019/02/08 閲覧】

[3] Digital Bibliography& Library Project: https://dblp.uni-trier.de/【2019/02/08 閲覧】