ISS-SP-039

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に関する検討

林　容央†　　桂井　麻里衣†
† 同志社大学理工学部インテリジェント情報工学科

1. はじめに

異分野融合や産学連携を試みるにあたり，招集する研究者を見つけることが難しい場面が多々存在する．個々の研究者の活動を把握するのは労力を要するため，こうした招集は人づての情報に依存している現状がある．この問題の解決策として，ウェブデータから研究者の業績，経歴，記事，講演情報や出版物などを自動集約し，そこから各研究者の活動情報を抽出して推薦に生かすことが考えられる．

本稿では，研究者のウェブページとそのカテゴリラベルからなるデータセットを構築し，ページのURL文字列とテキストに基づく多クラス分類手法を提案する．

2. 提案手法

2.1 データセット

はじめに，研究者名とその所属情報をKAKENデータベースからランダムに477名分収集した．次に，研究者名と大学名をGoogle検索のクエリとし，得られたウェブページのURLとテキストを保存したあと，表1に示す7種類のカテゴリラベルのいずれかに手動で分類した．スクレイピングできないページ（PDFファイルなど）や学術研究に使用できないサイト（researchmapなど）は除外した．データ数は5,267件となった．その中から，学習用データとテストデータへ9:1に分割した．

表1 研究者ウェブページのカテゴリ．

1	経歴, 業績	5	リポジトリ
2	出版物	6	記事
3	講演情報	7	その他
4	スタッフ一覧		

2.2 分類モデルの構築

提案手法では，ウェブページの本文（HTMLタグを除外したテキスト）の単語ID列における最初の128文字と，図1のように抽出されるURLのpage path（PP）テキストを入力として多クラス分類を行う．本文からの特徴抽出には，東北大学が公開している訓練済み日本語BERT [2]の最終層から得られる先頭のベクトルを全結合層へ入力し，分類問題にファインチューニングする．

PPテキストは，convolutional neural network（CNN）[4]とneural network（NN）を用いて特徴抽出する．CNNの畳み込み層では3つのカーネルを用いる．NNは，ニューロン数を100，隠れ層を8層とする．

本文と PP のそれぞれでネットワークをそれぞれ学習またはファインチューニングしたあと，BERT の最終層の先頭ベクトルと PP の予測結果ベクトルを全結合層へ入力し，最終的な予測のための結合モデルを構築する．図2に概要を示す．

- 「https://search.adb.fukushima-u.ac.jp/Profiles/2/0000164/profile.html」のオレンジ部分
- 「https://www.cse.sci.waseda.ac.jp/department/interview/sasakiyoh/」のオレンジ部分
- 「https://www.chitose.ac.jp/course/teacher/karthaus」のオレンジ部分

図 1 URL の PP テキストの抽出例．

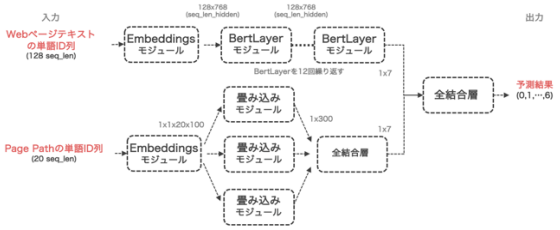

図 2 BERT+CNN 結合モデル．

表 2 テストデータにおける性能．

モデル	入力	Accuracy	Precision	Recall	F1
ランダム	-	0.145	0.147	0.147	0.139
NN	PP	0.396	0.496	0.403	0.428
CNN	PP	0.572	0.642	0.573	0.596
BERT	本文	0.694	0.731	0.682	0.700
BERT+NN	本文& PP	0.688	0.702	0.668	0.679
BERT+CNN	本文& PP	0.707	0.741	0.705	0.720

3. 実験

多クラス分類の評価指標として，Accuracy, Precision, Recall, macro-F1 スコアを用いる．BERT のファインチューニング，NN および CNN の学習はいずれも 50 エポックに固定し，macro-F1 が最も高いモデルでテストデータを分類した．その性能を表 2 に示す．URL の PP を入力としている NN と CNN では，CNN の精度が NN に比べて性能が大幅に高い．この結果は，単語埋め込みの空間情報を失わずに学習できる CNN の強みが表れていると考える．本文を入力としている BERT と結合すると，NN は悪影響を及ぼし，BERT 単体のスコアよりも低いスコアとなった．それに比べ，CNN と BERT を結合すると，本実験で一番高いスコアが得られた．

4. まとめ

本稿では，研究者氏名と所属名をクエリとした際の検索結果と，7 種類のカテゴリラベルからなるデータセットを構築した．また，サイトの本文に基づく BERT 特徴量と URL の PP に基づく CNN 特徴量から多クラス分類器を構築した．今後は，分類に使用するネットワークをさらに検討するほか，ウェブページのカテゴリを考慮した情報抽出に取り組む．

参考文献

[1] Chamoso, P. et al., “Profile generation system using artificial intelligence for information recovery and analysis.” Journal of Ambient Intelligence and Humanized Computing, 11 (2020): 4583–4592.

[2] Devlin, J. et al., "BERT: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[3] Gupta, A. and Rajesh, B. "Ensemble approach for web page classification." Multimedia Tools and Applications, 80(16) (2021): 25219–25240.

[4] Jia, Yangqing et al. "Caffe: Convolutional architecture for fast feature embedding." *Proceedings of the 22nd ACM international conference on Multimedia*, (2014): 675–678.